Iterative Re-Constrained Group Sparse Face Recognition With Adaptive Weights Learning

Article *in* IEEE Transactions on Image Processing · March 2017

DOI: 10.1109/TIP.2017.2681841

5 authors, including:

Jianwei Zheng
Zhejiang University of Technology
31 PUBLICATIONS **107** CITATIONS
SEE PROFILE

Ping Yang
Zhejiang University of Technology
4 PUBLICATIONS **28** CITATIONS
SEE PROFILE

Iterative Re-constrained Group Sparse Face Recognition with Adaptive Weights Learning

Jianwei Zheng, Ping Yang, Shengyong Chen, *Senior Member, IEEE*,Guojiang Shen and Wanliang Wang

***Abstract*—In this paper, we consider the robust face recognition problem via iterative re-constrained group sparse classifier with adaptive weights learning (IRGSC). Specifically, we propose a group sparse representation classification (GSRC) approach in which weighted features and groups are collaboratively adopted to encode more structure information and discriminative information than other regression based methods. In addition, we derive an efficient algorithm to optimize the proposed objective function, and theoretically prove the convergence. There are several appealing aspects associated with IRGSC. First, adaptively learned weights can be seamlessly incorporated into the GSRC framework. This integrates the locality structure of the data and validity information of the features into $l_{2,p}$-norm regularization to form a unified formulation. Second, IRGSC is very flexible to different size of training set as well as feature dimension thanks to the $l_{2,p}$-norm regularization. Third, the derived solution is proved to be a stationary point (globally optimal if $p\geq$1). Comprehensive experiments on representative datasets demonstrate that IRGSC is a robust discriminative classifier which significantly improves the performance and efficiency compared with the state-of-the-art methods in dealing with face occlusion, corruption, and illumination changes, etc.**

***Index Terms*—Sparse representation, classification, weights learning, Group constraints, face recognition**

I. Introduction

HIGH sample dimensionality and short insufficiency of prior knowledge about the valid features for classification are two challenging problems in machine learning and pattern recognition. Face recognition (FR) remains an active topic after comprehensive research over the last two decades not just owe to its great application potential [1], [2]. It offers a good testbench to reveal how these two key machine learning problems are solvable as large and unambiguous trial databases are available for FR problem. Appearance based method [3], also known as holistic method, which exploits machine learning approaches on the complete image for both feature selection and classification, provides a credible way to cope with these two difficult machine learning problems. However, robust FR w.r.t. occlusion/corruption is still an open problem due to the variations of noises, such as real disguise, continuous or pixel-wise occlusion, randomness of occlusion position, and the percentage of occluded pixels.

Recently, regression analysis based approaches arouse broad interests in FR community. Naseem et al. presented a linear regression classifier (LRC) for FR [4], which represents the query image by a linear combination of the class-specific dictionary atoms. Wright et al. proposed a sparse representation based classification (SRC) algorithm to identify face images with different corruption and real disguise [5]. In SRC, a query image is regressed as a sparse linear combination of the dictionary atoms from all classes collaboratively, and then the decision is made by identifying which subject yields the minimal reconstruction residual. Both LRC and SRC cannot achieve desirable performance especially when the dictionary is not overcomplete enough, which is common in practical FR system. Yang et al. [6] gave an insight into SRC and provided some theoretical supports for its effectiveness. They asserted that it is l_1 regularizer rather than l_0 (which counts the number of nonzero entries in a vector) that renders SRC resultful. Zhang et al. [7] analyzed the core theory of SRC and argued that the collaborative mechanism plays a more essential role than the l_1-norm based sparsity constraint. They proposed a collaborative representation classifier (CRC) based on l_2-norm constraint. By using of the structural information of errors, Yang et al. [8] further proposed nuclear norm based matrix regression (NMR) classification framework under l_2-norm regularization for better FR performance in the presence of occlusion and illumination variations. Zhong et al. believed that there should be a certain balance between SRC and CRC. They introduced $l_{1/2}$-norm regularization for the practical FR which claimed to embrace both advantages of SRC and CRC for better performance [9]. The label of the dictionary atoms is not utilized in the aforementioned algorithms, hence their regression is based solely on the structure of the individual samples. Nie et al. introduced the $l_{2,1}$-norm regularity term to incorporate the class labels. Their method, named as group sparse classifier (GSC), explores the sparse structure in the group form with more discriminative information [10].

It is worth mentioning that, these classifiers and their variants, however, still have not overcome the two elementary constraints of regression type algorithms. One is the carefully navigated importance of different samples and the other is the removal of some invalid features. A violation of these two conditions results in poor performance of the regression-based classification [11], [12]. The first constrain makes all the training samples equally contribute to discriminant, which is unrealistic. The second constrain renders regression based approaches perform poorly for the heavy corrupted query

This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 61602413, 61325019, U1509207, 61502424 and 61379123 and in part by the Zhejiang Provincial Natural Science Foundation under Grant LY15F030014. (Corresponding author: Shengyong Chen.)

J. Zheng, P. Yang, G. Shen and W. Wang are with the college of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China (e-mail: zjw@zjut.edu.cn; yizhongyangping@126.com; gjshen1975@zjut.edu.cn; wwl@zjut.edu.cn).

S. Chen is with the college of Computer and Communication Engineering, Tianjin University of Technology, Tianjin 300384, China, and also with the college of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China (e-mail: csy@zjut.edu.cn).

samples. Thus it is not a surprise that, despite of the impressive results of regression classifiers and various extensions, many works [11], [13]–[15] show doubts about their validity for image classification.

To overcome the first constraint of regression classifiers, query-adapted technique, also called distance weights learning [16], provides an idea to separate outlier samples from the dictionary atoms. A weights vector is obtained by computing the distances from query sample to the whole training samples [17]. Accordingly, many developed weighted regression algorithms benefit from this simple idea, including weighted SRC (WSRC) [18], weighted CRC (WCRC) [19], and locality group sensitive sparse representation (LGSR) [20]. Tang et al. argued that locality weighted regularized term of LGSR disrupts its group structure of sparse solution. Considering that each class plays a different role in regressing the query samples, they presented a weighted GSC (WGSC) [21] algorithm with consideration of the influence of the similarity between query samples and classes. However, the distribution structure of the training samples may not coincide with the natural class structure for a FR problem due to the influence of corruption. As a result, in seeking a weights vector, these classifiers may undesirably remove some samples which are in fact needed to represent the query image. Timofte et al. [12] adopted fixed point theorem to fully exploit the weights information from dictionary atoms, without employing query-adapted technique. They claimed that their proposed method has higher computational efficiency than WCRC and WSRC, while it keeps the same recognition performance. However, their method requires carefully controlled training images with both quality and quantity, which is difficult to achieve in practice.

There are some attempts to alleviate the second constraint of regression classifiers. RSRC [5] introduces an identity matrix as a dictionary to code the outlier features (e.g., pixels with corruption or occlusion). Naseem et al. [22] and Zhang et al. [23] respectively extended their LRC and CRC to the robust version, robust linear regression classification (RLRC) and robust collaborative representation classification (RCRC), using the Huber and Laplacian estimator to deal with severe random pixel noise and illumination changes. To unify the existing robust sparse regression models: the additive model for error correction and multiplicative model for error detection, He et al. [24], [25] created a half-quadratic framework by defining different half-quadratic functions based on the maximum correntropy criterion. Borrowing the idea of matrix based representation from NMR, Luo et al. [26] and Chen et al. [27] respectively introduced matrix variate slash and elliptically contoured distribution to image representation for better noise resistant. In addition, Yang et al. sought for a maximum a posterior solution and proposed a regularized robust coding (RRC) model for FR [28], which is robust to various types of feature outliers (e.g. corruption and facial expression). Qian et al. [29] further extended RRC to robust general regression and representation model (RGRR) by using of the prior information of the training set. RGRR works well when the query samples share the same probability distribution with the training samples. However, it involves an independent training stage that is unnecessary for the traditional regression based classification approaches. To sum up, although much progress has been made, robust FR is still an open issue due to the complex variation of corruption.

In this work, we manage to solve the two elementary limitations of regression based classification methods. We propose an iterative re-constrained group sparse classification (IRGSC) to increase the robustness of FR in dealing with severe occlusion, complex corruption, real disguises and large expression variation. The main contributions of this paper are outlined as follows:

1) A general framework is presented for regression-based classification. It unifies previous l_1, l_2 or $l_{2,1}$ regularized norm into a general formulation and learns the feature weights and distance weights simultaneously to achieve the optimal representation coefficients. We derive a new and efficient algorithm to iteratively and adaptively update the weights vector. Especially for the feature weights, we present a closed solution seamlessly connected to the model with only one univocal parameter, while the existing approaches all rely on different distribution functions for corresponding noises.

2) We extend the convex group norm to a concave surrogate function for a tighter approximation of the $l_{2,0}$-norm. Then the weighted $l_{2,p}$-norm penalty is enforced on the coefficients for purpose of imposing both distance locality and group sparsity, where p is released from a fixed value and is flexible to various training size and feature dimension. By introducing the feature weights vector to compute the reconstruction residuals and distance measurement, the proposed approach uses selected features to reflect the true distribution structure. Compared with WGSC, the sparse solution of IRGSC maintains locality at a feature level and contains more discriminative information.

3) In our implementation, the IRGSC minimization problem is transformed into an iteratively re-constrained group sparse coding problem with a reasonably designed weight learning strategy for robust FR. In theory, we prove that IRGSC monotonically decreases the objective value and any coefficients sequence is a stationary point. Our extensive experiments in benchmark face databases show that IRGSC achieves much better performance than existing regression based FR classifiers, especially when there are complicated variations, such as severe occlusions and corruptions, etc.

The rest of this paper is organized as follows: we introduce a general formulation of regression based classifiers in Section 2. Section 3 introduces IRGSC classifier for face recognition. In Section 4, we present the optimization algorithm of IRGSC. Section 5 analyses the complexity and convergence of the proposed method. In Section 6, we conduct experiments on 3 public face databases and compare our results with the state-of-the-art methods. Finally, Section 7 concludes the paper.

II. A General Framework for Regression based Classifier

For various classification task, different regression based approaches have been proposed due to the variation of the motivations, however, their purposes are often similar in the sense that they aim to derive a series of representation

coefficients and facilitate the succeeding classification task. A natural question that arises is whether these methods can be reformulated into a unifying framework and whether this framework assists in deriving new classifiers. In this section, we give positive answers to this question. We present a unified formulation of regression representation to provide a common perspective in comprehending the relationship among the existed algorithms and to devise new classifiers.

For a general classification problem, the training samples are represented as a dictionary matrix $\boldsymbol{X}=[\boldsymbol{X}_1, \boldsymbol{X}_2,..., \boldsymbol{X}_c]\in \mathbf{R}^{m\times n}$ supposing there exist c classes of subjects, where $\boldsymbol{X}_i=[\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2},..., \boldsymbol{x}_{in_i}]\in\mathbf{R}^{m\times n_i}$ is the sample subset from subject i=1,2,...,c. $\boldsymbol{x}_{ij}\in\mathbf{R}^m$ is the jth sample from the ith class with feature dimension m. Here n_i is the number of training samples of class i, and $n=\sum_{i=1}^{c} n_i$ is the total sample number. The aim of representation based classification is, given the training set $\boldsymbol{X}$, to correctly determine the class to which a query test sample $\boldsymbol{y}\in\mathbf{R}^m$ belongs.

For this purpose, regression methods use the dictionary $\boldsymbol{X}$ to represent the query $\boldsymbol{y}$ linearly as

$$\begin{aligned}\boldsymbol{y} &= \boldsymbol{X}_1\boldsymbol{\theta}_1 + \boldsymbol{X}_2\boldsymbol{\theta}_2 + ... + \boldsymbol{X}_c\boldsymbol{\theta}_c \\ &= \boldsymbol{x}_{11}\boldsymbol{\theta}_{11} + \boldsymbol{x}_{12}\boldsymbol{\theta}_{12} + ... + \boldsymbol{x}_{cn_c}\boldsymbol{\theta}_{cn_c} \\ &= \boldsymbol{X}\boldsymbol{\theta},\end{aligned} \tag{1}$$

where $\boldsymbol{\theta}=[\theta_{11},\theta_{12},...,\theta_{cn_c}]^{\mathrm{T}}\in\mathbf{R}^n$ is the coefficient vector to be determined. Suppose the optimal representation vector $\boldsymbol{\theta}^*$ is achieved, let $\delta_i(\boldsymbol{\theta}^*)$ be the vector whose only nonzero entries are the entries of $\boldsymbol{\theta}^*$ associated with class i. The query $\boldsymbol{y}$ can be reconstructed by the training samples of class i as $\boldsymbol{y}_i=\boldsymbol{X}\delta_i(\boldsymbol{\theta}^*)$, i=1,...,c. The label of $\boldsymbol{y}$ is decided as the class which gives the minimum reconstruction error [28]

$$\mathrm{identity}(\boldsymbol{y}) = \arg\min_i \|\boldsymbol{y} - \boldsymbol{X}\delta_i(\boldsymbol{\theta}^*)\|. \tag{2}$$

By surveying the existing algorithms, we introduce the following criteria to uniformly compute the representation coefficients

$$\min_{\theta} \|\boldsymbol{s}\odot(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})\|_q^q + \lambda\|\boldsymbol{\eta}\odot\boldsymbol{\theta}\|_p^p, \tag{3}$$

where $\odot$ denotes element-wise multiplication, $\|\boldsymbol{u}\|_x$ is the l_x-norm of $\boldsymbol{u}$. The objective function (3) consists of two parts: the first part measures the reconstruction error while the second one is a regularization term, with λ representing the regularization parameter that balances the contribution of the reconstruction error and locality of the solution. In the following, we give an overview of these regression type algorithms with various implementation details of p, q, $\boldsymbol{s}$ and $\boldsymbol{\eta}$.

A. Linear Regression type Classifier

The traditional linear representation-type algorithms take no considerations of locality or feature weight factor. That is to say, they let both of $\boldsymbol{s}$ and $\boldsymbol{\eta}$ to be $\boldsymbol{1}$ (a vector with all entries be 1) in (3) and only employ different forms of norm constraints to characterize the representation coefficients. SRC [5] deems that sparsity plays the most important role in capturing discriminative information. Hence the authors employ l_1-norm (i.e.,$\|\boldsymbol{\theta}\|_1=\sum_{i=1}^{n}|\theta_i|$) to recover the solution of Eq.(1) as follows

$$\min_{\theta} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1. \tag{4}$$

However, some researchers are skeptical about whether the sparsity constraint is necessary in classification. They argue that the success of SRC is attributed to collaborative mechanism rather than sparsity. Consequently, CRC [7] is proposed, which replaces the l_1-norm in (4) with l_2-norm (i.e.,$\|\boldsymbol{\theta}\|_2=\sum_{i=1}^{n}\theta_i^2$). On the other hand, some researchers doubt that in practical cases, the l_1-norm cannot achieve solutions as sparse as l_0-norm since the dictionary is not overcomplete enough. Since the combinatorial l_0-norm minimization is an NP-hard problem, the $l_{1/2}$-norm (i.e.,$\|\boldsymbol{\theta}\|_{1/2}=\sum_{i=1}^{n}|\theta_i|^{1/2}$) minimization, as a closer constraint to l_0-norm than l_1-norm, is employed in LHC [9] for sparse coding.

Another limitation of SRC is that the label information of each training sample is not considered when solving the l_1-norm minimization problem. Therefore, SRC might represent a test sample by training samples from unrelated classes, and thus is not preferable for classification. GSC [10] tries to find a coefficient $\boldsymbol{\theta}$ which is sparse at group level. This means the non-zero coefficients of $\boldsymbol{\theta}$ just occur at few specific groups, and meanwhile the coefficients within the selected groups are non-sparse. GSC seeks for a representation that uses the minimum number of groups instead of atoms. Hence, the $l_{2,1}$ mixed norm (i.e.,$\sum_{i=1}^{c}\|\theta_i\|_2$) regularizer is used in the optimization process, where the query $\boldsymbol{y}$ is reconstructed by samples from few classes.

B. Weighted Regression type Classifier

The aforementioned classifiers emphasize that sparsity or collaborative mechanism is important in representing the query sample, however, they neglect the locality constraint, which is more important in revealing the true geometry of feature space [30], [31]. In other words, the query $\boldsymbol{y}$ might be represented by training samples that are far away from it. Many locality-constrained linear classifiers have been proposed recently. Specifically, they integrate $\boldsymbol{\eta}$ into norm constraints in order to avoid selecting the training samples that are far from $\boldsymbol{y}$ to represent the test sample

$$\min_{\theta} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\eta}\odot\boldsymbol{\theta}\|_p^p, \tag{5}$$

where $\boldsymbol{s}=\boldsymbol{1}$ and q=2 are fixed, $\boldsymbol{\eta}=[\eta_{11},\eta_{12},...,\eta_{cn_c}]^{\mathrm{T}}$ measures the similarity between the query sample and all the reference samples. Specifically,

$$\eta_{ij} = \exp(\|\boldsymbol{x}_{ij} - \boldsymbol{y}\|_2^2/\sigma^2), i = 1, ...c, j = 1, ..., n_i, \tag{6}$$

where σ is the bandwidth parameter and η_{ij} denotes the distance between $\boldsymbol{y}$ and $\boldsymbol{x}_{ij}$. Clearly, the larger the η_{ij} is, the smaller the weight coefficient is. In (5), when p is assigned as 1 or 2, the corresponding weighted classifier is WSRC [18] and WCRC [19], respectively.

Similar to WSRC and WCRC, LGSR [20] combines the reconstruction error with data locality constrained group sparsity regularizer. However, the locality constraint regularized

term disrupts the group structure of sparse solution. WGSC [21] imposes not only locality constraints on group sparsity to exclude the training samples which are far away from the test sample, but also considers the similarity between the test sample and each class. Accordingly, the weight coefficients of WGSC is formulated as $\boldsymbol{\eta}=\boldsymbol{r}\odot\boldsymbol{d}$, where $\boldsymbol{d}$ is computed same as (6), and $\boldsymbol{r}$ is used to assess the relative importance of all the reference classes. Inspired by LRC [9], $\boldsymbol{r}$ is computed as

$$r_i = \|\boldsymbol{y} - \boldsymbol{X}_i\boldsymbol{\theta}_i^*\|_2^2, \boldsymbol{\theta}_i^* = \arg\min \|\boldsymbol{y} - \boldsymbol{X}_i\boldsymbol{\theta}_i\|_2^2. \quad (7)$$

C. Robust Regression type Classifier

Although the diversity of norm constrained regularization, the aforementioned classifiers all employ the l_2-norm to characterize the reconstruction residual. Considering the fact that the l_2-norm is sensitive to large outliers, RSRC [5] adopts l_1-norm into SRC for prompting the residual estimation to be more robust. Similarly, RCRC [23] characterizes the representation fidelity of CRC with l_1-norm for robustness to corruptions/occlusions. That is to say, the value of q in (4) is set to be 1. From the perspective of maximum likelihood estimation, the l_2-norm and l_1-norm are based on the hypothesis that the error residuals are independent identically distributed with Gaussian distribution or Laplacian distribution. RRC [28] models sparse coding as a robust regression problem and adopts an adaptive distribution to characterize the noises. Specifically, RRC uses the following criteria

$$\min_{\theta} \|\boldsymbol{s} \odot (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})\|_2^2 + \lambda\|\boldsymbol{\theta}\|_p^p, \quad (8)$$

where p can be 1 or 2 (termed as RRCL1 or RRCL2, respectively). Rewrite $\boldsymbol{X}$ as $\boldsymbol{X}=[\boldsymbol{z}_1;\boldsymbol{z}_2;...;\boldsymbol{z}_m]$, where $\boldsymbol{z}_i \in \mathbf{R}^n$ is the ith row of $\boldsymbol{X}$, and let $\boldsymbol{e}=\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta}=[e_1;e_2;...;e_n]$, where $e_i=y_i-\boldsymbol{z}_i\boldsymbol{\theta}$, i=1,2,...,m. The feature weight $\boldsymbol{s}$ in (8) is set to be the following logistic function

$$s_i = \frac{\exp\left(-\mu e_i^2 + \mu\delta\right)}{1 + \exp\left(-\mu e_i^2 + \mu\delta\right)}, \quad (9)$$

where μ and δ are positive scalars. Parameter μ manipulates the decreasing rate from 1 to 0, and parameter δ determines the location of demarcation point. With the help of $\boldsymbol{s}$, RRC prefers to assign larger weights to inliers and smaller weights to outliers; that is, it has higher capability to classify inliers and outliers.

III. Iterative Re-constrained Group Sparse Classification

In addition to encompassing most popular regression type classification algorithms, the criteria (3) can also serve as a general framework under which new algorithms for classification to be derived. The mentioned works in Section 2 only employ limited advantages to their models. For example, SRC just emphasizes that strong sparsity will bring about strong discriminant power. RRC assigns different weights to different features according to their coding residuals, but without consideration of locality. Following the intuition that group sparsity, distance locality, and weighted features explicitly encourages a better representation-type classifier, we incorporate all these advantages into the objective function. Correspondingly, our preliminary objective is

$$\min_{\theta} \|\boldsymbol{s} \odot (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})\|_2^2 + \lambda \sum_{i=1}^{c} \|\boldsymbol{\eta}_i \odot \boldsymbol{\theta}_i\|_2^p, \quad (10)$$

which consists of two parts: $\|s\odot(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta})\|$ is the reconstruction error with integration of feature weights $\boldsymbol{s}$, and the second part represents the weighted $l_{2,p}$ mixed norm based regularizer on the coefficient $\boldsymbol{\theta}$, where $p>0$.

A. Adaptive Feature Weights Learning

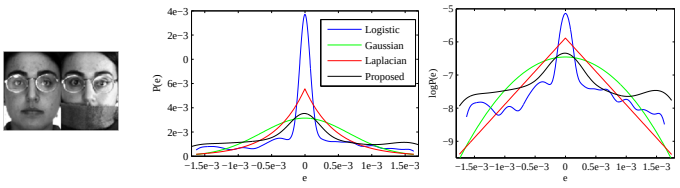
As indicated by [25], different $\boldsymbol{s}$ leads to different classifiers. When $\boldsymbol{s}$ is set to be 1 constantly, it corresponds to the l_2-norm fidelity as in CRC; when set as $s_i=1/|e_i|$, it turns to the l_1-norm fidelity as in SRC; when set to a logistic function as Eq.(9), it becomes RRC. However, all these functions fundamentally have certain limits. The l_2-norm fidelity treats all features equally, no matter whether they are outlier or not. The l_1-norm fidelity assigns infinity to features when the residual approaches to zero, making the coding unstable. The logistic weight function has two tunable parameters need to be set manually, which costs a lot of time. Moreover, the essential relationship between the logistic function and the real weights has not been theoretically revealed.

In this paper, we expect $\boldsymbol{s}$ to be more flexible, which is adaptive to the query $\boldsymbol{y}$ so that the classifier is more robust to mixed types of noises. We first use an example to illustrate different distributions of residual $\boldsymbol{e}$ by different models. Fig.1(a) is a clean face sample from the AR database, while Fig.1(b) presents the real disguised query face image $\boldsymbol{y}$ with scarf. The distributions of $\boldsymbol{e}$ by using Gaussian (SRC), Laplacian (RSRC), and the logistic function (RRC) are plotted in Fig.1(c). Fig.1(d) further illustrates the distributions in log domain for better observation of the tails. It has been reported repeatedly that the empirical distribution of $\boldsymbol{e}$ has a strong peak at zero but with a long tail, which is mostly caused by the occluded and corrupted features [32], [33]. From Fig.1(d), it can be seen that the distribution of logistic function has explicitly heavier tail than the Gaussian and Laplacian models, which explains why RRC works better than SRC and RSRC in handling occlusion and corruption. Moreover, it is well-known that for robust classifier, a well fitted tail is much more important than the fitting of the peak, which is generated by the small trivial coding residuals [25], [34], [35]. However, it can be seen that the distribution of RRC has a stronger peak and a lighter tail than the proposed method.

For achieving the goal delineated in Fig.1, an intuitive idea to determine the weights s_i, i=1,..., m is solving the following problem

$$\min_{s^{\mathrm{T}}\boldsymbol{I}=1, s_i\geq 0} \sum_{i=1}^{m} \|\boldsymbol{y}_i - \boldsymbol{z}_i\boldsymbol{\theta}\|_2^2 s_i, \quad (11)$$

which assigns smaller positive value to larger e_i. However, the criteria (11) has a trivial solution, only the smallest residual has corresponding s_i with value 1 and the other features are all having weights zero. On the other hand, if we solve the



(a) (b) (c) (d)

Fig. 1. The residual distributions of occluded face image by competing methods.(a) Clean image.(b) Occluded query image.(c) and (d) Distributions of coding residuals in linear and log domains.

following criteria without involving any residual information

$$\min_{s^{\mathrm{T}}\boldsymbol{I}=1,s_i\geq 0}\sum_{i=1}^{m}s_i^2, \tag{12}$$

the optimal solution is that all the feature weights equals to $1/m$ concurrently , which can be seen as the Gaussian prior [36]. Combining (11) and (12), we get the following criteria

$$\min_{s^{\mathrm{T}}\boldsymbol{I}=1,s_i\geq 0}\sum_{i=1}^{m}(\|\boldsymbol{y}_i-\boldsymbol{z}_i\boldsymbol{\theta}\|_2^2 s_i+\gamma s_i^2). \tag{13}$$

The second term in (13) is a regularization and γ is a tunable parameter. After some deduction, (13) can be transformed into

$$\min_{s^{\mathrm{T}}\boldsymbol{I}=1,s_i\geq 0}\|\boldsymbol{s}^{1/2}\odot(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta})\|_2^2+\gamma\|s\|_2^2. \tag{14}$$

Note that the l_1-norm $\|\boldsymbol{s}\|_1$ is also a commonly used regularization [5], [27] for non-trivial solution, but interestingly the constraints in problem (14) makes the l_1 term constant. Moreover, We will see in Subsection 3.3 that this problem can be solved with a closed form solution under the l_2-norm constraints.

B. Adaptive Distance Weights Learning

As we known, WGSC takes advantage of both group sparsity and local smooth sparsity constraints. However, the limitation of WGSC is that the features from the query sample are jointly selected. This may not be always preferable since different types of noises may occur. To deal with this problem, we impose feature constraints on distance weights to adaptively exclude the pixels which are far away from the true subject. Similar to WGSC, the aim of our regression is that we try to represent the query $\boldsymbol{y}$ by training samples which are not only from the neighbors, but also from the highly relevant classes. Hence, a weighted group sparsity representation method with feature constraints is proposed, which can be formulated mathematically as follows

$$\min_{s^{\mathrm{T}}\boldsymbol{I}=1,s_i\geq 0}\|\boldsymbol{s}^{1/2}\odot(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta})\|_2^2+\gamma\|\boldsymbol{s}\|_2^2+\lambda\sum_{i=1}^{c}r_i\|\boldsymbol{d}_i\odot\boldsymbol{\theta}_i\|, \tag{15}$$

where r_i is utilized to assess the relative importance of observations per class for representing the query sample. Inspired by WGSC, we set it as

$$\begin{aligned} r_i &= \|\boldsymbol{s}^{1/2}\odot(\boldsymbol{y}-\boldsymbol{X}_i\boldsymbol{\theta}_i^*)\|_2,\\ \boldsymbol{\theta}_i^* &= \arg\min_{\boldsymbol{\theta}_i}\|\boldsymbol{s}^{1/2}\odot(\boldsymbol{y}-\boldsymbol{X}_i\boldsymbol{\theta}_i^*)\|_2^2\\ &= (\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{S}\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{S}\boldsymbol{y}, \end{aligned} \tag{16}$$

where $\boldsymbol{S}$=diag([$\boldsymbol{s}_1$,$\boldsymbol{s}_2$,...,$\boldsymbol{s}_c$])$\in\mathrm{R}^{m\times m}$ is a diagonal matrix. A large r_i would drive the matching coefficients of ith subject, $\boldsymbol{\theta}_i$, shrink to zeros. Moreover, for eliminating outliers, we further integrate locality constraints $\boldsymbol{d}_i$ into the group norm. In (15), $\boldsymbol{d}_i$=[d_{i1},d_{i2},...,d_{in_i}]$^{\mathrm{T}}\in\mathbf{R}^{n_i\times 1}$ penalizes the Euclidean distance between query $\boldsymbol{y}$ and the training samples. The definition of d_{ik}, k=1,...,n_i, is

$$d_{ik}=\|\boldsymbol{s}^{1/2}\odot(\boldsymbol{y}-\boldsymbol{x}_{ik})\|_2^2. \tag{17}$$

In this manner, to precisely represent a query sample, we not only consider the group similarity and data locality, but also take into account feature contributions prompted by mixed types of noises. Further denote η_i=[$\eta_{i1},\eta_{i2},...,\eta_{in_i}$]$^{\mathrm{T}}$, i=1,...,c, where η_{ik}=$r_i d_{ik}$. Then our objective criteria can be reformulated as

$$\begin{aligned} E(\boldsymbol{\theta},\boldsymbol{s}) = &\min_{s^{\mathrm{T}}\boldsymbol{I}=1,s_i\geq 0}\|\boldsymbol{s}^{1/2}\odot(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta})\|_2^2\\ &+\gamma\|\boldsymbol{s}\|_2^2+\lambda\sum_{i=1}^{c}\|[\eta_i\odot\boldsymbol{\theta}_i]\|_2^p. \end{aligned} \tag{18}$$

The major challenge for solving (18) is that the latter term of the objective function is non-smooth. Inspired by [37], it can be smoothed by introducing the following regularization,

$$\begin{aligned} E(\boldsymbol{\theta},\boldsymbol{s},\mu) = &\min_{s^{\mathrm{T}}\boldsymbol{I}=1,s_i\geq 0}\|\boldsymbol{s}^{1/2}\odot(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta})\|_2^2\\ &+\gamma\|\boldsymbol{s}\|_2^2+\lambda\sum_{i=1}^{c}\|\begin{bmatrix}\eta_i\odot\boldsymbol{\theta}_i\\ \mu\end{bmatrix}\|_2^p, \end{aligned} \tag{19}$$

where $\mu>0$ is a constant scalar used to make the objective function smooth (see Eq.(28)). Replacing the non-smoothed problem Eq.(18) with Eq.(19) brings several advantages. First, a smooth objective function usually makes the optimization easier. Second, if $p\geq 1$, the problem of $E(\boldsymbol{\theta},\boldsymbol{s},\mu)$ is convex, this guarantees a globally optimal solution.

Theorem 1: If $p\geq 1$, $E(\boldsymbol{\theta},\boldsymbol{s},\mu)$ is convex w.r.t $\boldsymbol{\theta},\boldsymbol{s}$ and μ. Also, for any $\mu>0$, $E(\boldsymbol{\theta},\boldsymbol{s})$ is convex w.r.t $\boldsymbol{\theta}$ and $\boldsymbol{s}$.

By adopting the convexity of l_2-norm and $l_{2,p}$-norm when $p\geq 1$, the above theorem can be easily proved.

Third, it is clear that $E(\boldsymbol{\theta},\mu)\geq E(\boldsymbol{\theta})$, where the equality holds if and only if μ=0. Indeed, by the fact that

$$\begin{aligned} \sum_{i=1}^{c}\|\begin{bmatrix}\eta_i\odot\boldsymbol{\theta}_i\\ \mu\end{bmatrix}\|_2^p &= \sum_{i=1}^{c}((\eta_i\odot\boldsymbol{\theta}_i)^{\mathrm{T}}(\eta_i\odot\boldsymbol{\theta}_i)+\mu^2)^{p/2}\\ &\geq \sum_{i=1}^{c}\|\eta_i\odot\boldsymbol{\theta}_i\|_2^p. \end{aligned}$$

That is to say, Eq.(19) is majorized by Eq.(18) with any given μ. Decreasing Eq.(19) tends to decrease Eq.(18). Moreover, for any $\varepsilon>0$, there exists $\mu>0$ satisfying $E(\boldsymbol{\theta},\mu)\leq E(\boldsymbol{\theta})+\varepsilon$. Suppose that $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}'^*$ are the optimal solutions to (19) and

(18), respectively. We have

$$0 \leq E(\boldsymbol{\theta}^*) - E(\boldsymbol{\theta}'^*) \leq E(\boldsymbol{\theta}'^*) - E(\boldsymbol{\theta}'^*) + \varepsilon = \varepsilon.$$

We would say that the solution $\boldsymbol{\theta}^*$ to (19) is ε-optimal w.r.t.(18).

IV. Algorithm of the Proposed Model

As illustrated in Section 3, the minimization of $E(\boldsymbol{\theta}, \boldsymbol{s}, \mu)$ is an iterative process, and the weights $\boldsymbol{S}$ and $\boldsymbol{\eta}$ are updated sequentially for the expected coding coefficients $\boldsymbol{\theta}^*$.

When $\boldsymbol{\theta}$ is fixed, the problem (19) becomes

$$\begin{aligned} &\min_{s} \|\boldsymbol{s}^{1/2} \odot (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})\|_2^2 + \gamma \|\boldsymbol{s}\|_2^2 \\ &\quad s.t. \;\; \boldsymbol{s}^{\mathrm{T}}\boldsymbol{1} = 1, s_i \geq 0, i = 1 \sim m. \end{aligned} \tag{20}$$

Since e_i=y_i-$\boldsymbol{z}_i\boldsymbol{\theta}$, i=1,...,m are the entries of $\boldsymbol{e}\in\mathbf{R}^{m\times 1}$, problem (20) can be rewritten as

$$\begin{aligned} &\min_{s} \sum_{i=1}^{m} \{s_i e_i^2 + \gamma s_i^2\} = \min_{s} \|\boldsymbol{s} + \frac{\boldsymbol{e}}{2\gamma}\|_2^2 \\ &\quad s.t. \;\; \boldsymbol{s}^{\mathrm{T}}\boldsymbol{1} = 1, s_i \geq 0, i = 1 \sim m. \end{aligned} \tag{21}$$

The Lagrangian function of problem (21) is

$$L(\boldsymbol{s}, \kappa, \boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{s} + \frac{\boldsymbol{e}}{2\gamma}\|_2^2 - \kappa(\boldsymbol{s}^{\mathrm{T}}\boldsymbol{1} - 1) - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{s}, \tag{22}$$

where κ and $\boldsymbol{\beta}\geq 0$ are the Lagrangian multipliers. According to the KKT condition [38], we can get the optimal solution $\boldsymbol{s}$ as

$$\boldsymbol{s} = (-\frac{\boldsymbol{e}}{2\gamma} + \kappa)_+. \tag{23}$$

In practical scenarios, we could achieve better performance if we select limited numbers of features due to the impact of noises. Therefore, it is preferred to learn a spare $\boldsymbol{s}$ with parts of its features be 0. Another benefit of learning a sparse vector $\boldsymbol{s}$ is that the computational cost will be alleviated to an extent for subsequent processing.

Without loss of generality, suppose e_1,..., e_m are ordered from small to large. If the optimal $\boldsymbol{s}$ has l>0 zero elements, then according to Eq.(23), we get s_{m-l}>0 and s_{m-l+1}=0, that is

$$\begin{cases} -\frac{e_{m-l}}{2\gamma} + \kappa > 0 \\ -\frac{e_{m-l+1}}{2\gamma} + \kappa = 0 \end{cases} \tag{24}$$

Furthermore, according to the constraint $\boldsymbol{s}^{\mathrm{T}}\boldsymbol{1}$=1, we have

$$\begin{aligned} &\sum_{j=1}^{m-l}(-\frac{e_j}{2\gamma} + \kappa) = 1 \\ &\Rightarrow \kappa = \frac{1}{m-l} + \sum_{j=1}^{m-l} \frac{e_j}{2\gamma(m-l)}. \end{aligned} \tag{25}$$

According to (24) and (25), for obtaining an optimal solution $\boldsymbol{s}$ with exact l zero elements, we can set γ to be

$$\gamma = (m-l)\frac{e_{m-l+1}}{2} - \frac{1}{2}\sum_{j=1}^{m-l} e_j. \tag{26}$$

With derived parameters κ and γ, the optimal $\boldsymbol{s}$ can be get in a close-set form

$$s = (e_{m-l} - \boldsymbol{e})/((m-l)e_{m-l+1} - \sum_{j=1}^{m-l} e_j). \tag{27}$$

Note that the number of zero elements l is much easier to tune than the parameters μ and δ in RRC since l is an integer and has explicit sense.

By the fact that $\|\boldsymbol{z}\|_2^p$= $(\boldsymbol{z}^{\mathrm{T}}\boldsymbol{z})^{p/2}$, when $\boldsymbol{s}$ is fixed, the problem (19) becomes

$$\begin{aligned} f(\boldsymbol{\theta}) =& \min_{\theta} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^{\mathrm{T}}\boldsymbol{S}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}) \\ &+\lambda \sum_{i=1}^{c} ((\boldsymbol{\eta}_i \odot \boldsymbol{\theta}_i)^{\mathrm{T}}(\boldsymbol{\eta}_i \odot \boldsymbol{\theta}_i) + \mu^2)^{p/2}, \end{aligned} \tag{28}$$

For simplicity, let us denote $\boldsymbol{\Pi}$=diag([$\boldsymbol{\eta}_1$;$\boldsymbol{\eta}_2$;...;$\boldsymbol{\eta}_c$])$\in\mathbf{R}^{n\times n}$, and $\boldsymbol{\alpha}$=$\boldsymbol{\Pi}\boldsymbol{\theta}$, hence we have

$$\begin{aligned} E(\boldsymbol{\alpha}) =& \min_{\alpha} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\Pi}^{-1}\boldsymbol{\alpha})^{\mathrm{T}}\boldsymbol{S}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\Pi}^{-1}\boldsymbol{\alpha}) \\ &+\lambda \sum_{i=1}^{c} (\boldsymbol{\alpha}_i^{\mathrm{T}}\boldsymbol{\alpha}_i + \mu^2)^{p/2} \\ =& \min_{\alpha} (\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{\alpha})^{\mathrm{T}}\boldsymbol{S}(\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{\alpha}) \\ &+\lambda \sum_{i=1}^{c} (\boldsymbol{\alpha}_i^{\mathrm{T}}\boldsymbol{\alpha}_i + \mu^2)^{p/2}, \end{aligned} \tag{29}$$

where $\boldsymbol{X}'$=$\boldsymbol{X}\boldsymbol{\Pi}^{-1}$ can be computed in advance since $\boldsymbol{\Pi}$ is independent of $\boldsymbol{\alpha}$.

Theorem 2: $\boldsymbol{\theta}^*$=$\boldsymbol{\Pi}^{-1}\boldsymbol{\alpha}^*$, where $\boldsymbol{\alpha}^*$=$\arg\min_{\boldsymbol{\alpha}} E(\boldsymbol{\alpha})$, and $\boldsymbol{\theta}^*$=$\arg\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$.

Proof. From (28) and (29), we clearly have the equation $f(\boldsymbol{\theta})$=$E(\boldsymbol{\Pi}\boldsymbol{\theta})$. If we known $\boldsymbol{\alpha}^*$, $\forall\boldsymbol{\theta}$, $f(\boldsymbol{\theta})$=$E(\boldsymbol{\Pi}\boldsymbol{\theta})\geq E(\boldsymbol{\alpha}^*)$ $=f(\boldsymbol{\Pi}^{-1}\boldsymbol{\alpha}^*)$. Hence, $\boldsymbol{\theta}^*$=$\boldsymbol{\Pi}^{-1}\boldsymbol{\alpha}^*$.

From theorem 2, it shows that the solution of problem (28) can be directly transformed to problem (29). By setting the derivative of $E(\boldsymbol{\alpha})$ with respect to $\boldsymbol{\alpha}$ to zero, we have

$$\frac{\partial(E(\boldsymbol{\alpha}))}{\partial\boldsymbol{\alpha}} = -2\boldsymbol{X}'^{\mathrm{T}}\boldsymbol{S}\boldsymbol{y} + 2\boldsymbol{X}'^{\mathrm{T}}\boldsymbol{S}\boldsymbol{X}'\boldsymbol{\alpha} + 2\lambda\boldsymbol{D}\boldsymbol{\alpha} = \boldsymbol{0}, \tag{30}$$

where $\boldsymbol{D}$ is a block diagonal matrix defined by

$$\begin{vmatrix} \frac{p}{2\|\boldsymbol{\alpha}_1;\mu\|_2^{2-p}}\boldsymbol{I}_{n_1} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \ddots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \frac{p}{2\|\boldsymbol{\alpha}_c;\mu\|_2^{2-p}}\boldsymbol{I}_{n_c} \end{vmatrix}$$

From this, we obtain

$$\boldsymbol{\alpha} = (\boldsymbol{X}'^{\mathrm{T}}\boldsymbol{S}\boldsymbol{X}' + \lambda\boldsymbol{D})^{-1}\boldsymbol{X}'^{\mathrm{T}}\boldsymbol{S}\boldsymbol{y}. \tag{31}$$

Notice that both $\boldsymbol{s}$ and $\boldsymbol{D}$ depend on $\boldsymbol{\alpha}$. They can be computed if $\boldsymbol{\alpha}$ is fixed. On the other hand, if matrix $\boldsymbol{D}$ and feature weights $\boldsymbol{s}$ are fixed, $\boldsymbol{\alpha}$ can be obtained by solving (31). This fact motivates us to solve (19) by iteratively updating $\boldsymbol{\alpha}$ and $\boldsymbol{s}$, $\boldsymbol{D}$. This optimization strategy is called iteratively re-constrained group sparse classification (IRGSC), which is shown in Algorithm 1.

Algorithm 1 Iteratively Re-constrained Group Sparse Classification

Input: Normalized query sample $\boldsymbol{y}$ with l_2-norm; Normalized data matrix $\boldsymbol{X}\in\mathbf{R}^{m\times n}$; Class labels $\boldsymbol{c}$; Initialized $\boldsymbol{\alpha}^1$; Parameters λ, l and p; Iteration t=1 and max iteration m_i=50;

Output: $\boldsymbol{\theta}$*

1. Compute residual $\boldsymbol{e}^t = \boldsymbol{y}$-$\boldsymbol{X\alpha}^t$,
2. Estimate feature weights as Eq.(27).
3. Calculate the diagonal matrices $\boldsymbol{S}$, $\boldsymbol{D}$ and $\boldsymbol{\Pi}$.
4. Update the sparse coding coefficients $\boldsymbol{\theta}^*=\boldsymbol{\Pi}^{-1}\boldsymbol{\alpha}^*$, where $\boldsymbol{\alpha}^*$ is obtained from Eq.(31).
5. Go back to step 1 until the condition of convergence ($\|E^{t+1}$-$E^t\|_2/\|E^t\|_2<\xi$, ξ is a small positive scale) is met, or the maximal number of iterations is reached.

While the proposed IRGSC algorithm has a similar structure to the previous IRLS methods [39]–[41], they are in fact significantly different. First, majority of the reweighted schemes are specific to the regularization term of representation coefficient, while Algorithm 1 focuses on the achievement of robust data fidelity with desirable feature weights. Second, although a few works such as RRC [28] also employ reweighted scheme to the data fidelity term, they neglect the locality regularization on the coding coefficients and their scheme of constraints is not closely related to the objective function in theory.

V. Convergence and Complexity of IRGSC

A. Convergence of IRGSC

Previous regression type classifiers minimize the sum of a weighted regularization term and squared loss, while we minimize the sum of two weighted terms. In this section, we provide a new convergence analysis on IRGSC for optimization.

The critical links of Algorithms 1 are step 2 and step 4. Since the optimal $\boldsymbol{s}$ can be obtained in a close-set form (Eq.27) and the Hessian of Eq.(22) is positive definite. That means each iteration the objective function of Eq.(19) decreases by the step 2 operation. Considering that the cost function of Eq.(19) is lower bounded (>0), Algorithm 1 will converge if the step 4 operation will converge to a stationary point. For this purpose, we first introduce lemma 1 and prove the convergence of step 4.

Lemma 1: Given any nonzero vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, the following inequality holds for $p\in(0,\ 1]$:

$$\sum_{i=1}^{c}(\|\boldsymbol{y}_i\|_2^p - \|\boldsymbol{x}_i\|_2^p) \geq \mathrm{tr}((\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})\boldsymbol{M}),$$

where $\boldsymbol{M}$ is a diagonal matrix with the ith diagonal entry being $M_{ii} = \frac{p}{2}(\|\boldsymbol{y}_i\|_2^2)^{p/2-1}$.

Proof: Our proofs are grounded on the fact that $f(x)$=x^a is concave on x>0 when a$\in$(0, 1]. Given the nature of concave function, for any x, y>0, inequality y^a-x^a+ay^{a-1}(x-y)$\geq$0 holds, then we have

$$\begin{aligned}\|\boldsymbol{y}_i\|_2^p - \|\boldsymbol{x}_i\|_2^p \geq& \nabla\{(\|\boldsymbol{y}_i\|_2^p)\}(\|\boldsymbol{y}_i\|_2^2 - \|\boldsymbol{x}_i\|_2^2)\\ =&\frac{p}{2}(\|\boldsymbol{y}_i\|_2^2))^{(p/2-1)}(\boldsymbol{y}_i^{\mathrm{T}}\boldsymbol{y}_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_i).\end{aligned}$$

Summing all the above inequalities for all i=1,...,c, the proof of lemma 1 is completed.

Based on Lemma 1, we have the following theorems of Algorithm 1.

Theorem 3: The sequence $\{\boldsymbol{\alpha}^t\}$ generated in step 4 of Algorithm 1 possesses the following properties:

(1) $E(\boldsymbol{\alpha}^{t+1})\leq E(\boldsymbol{\alpha}^t)$;

(2) Sequence $\{\alpha^t\}$ is bounded;

(3) $\lim_{t\to\infty}\|\boldsymbol{\alpha}^t$-$\boldsymbol{\alpha}^{t+1}\|_2 = 0$.

Proof: Denote $\boldsymbol{e}_s^t$=$\boldsymbol{s}^{1/2}\odot$($\boldsymbol{y}$-$\boldsymbol{X\alpha}^t$). Since $\boldsymbol{\alpha}^{t+1}$ solves Eq.(30), then

$$\lambda\boldsymbol{D}^t\boldsymbol{\alpha}^{t+1} = \boldsymbol{X}'^{\mathrm{T}}\boldsymbol{S}\boldsymbol{y} - \boldsymbol{X}'^{\mathrm{T}}\boldsymbol{S}\boldsymbol{X}'\boldsymbol{\alpha}^{t+1}. \tag{32}$$

A dot product with $\boldsymbol{\alpha}^t$-$\boldsymbol{\alpha}^{t+1}$on both side of Eq.(32) gives

$$\begin{aligned}&\lambda(\boldsymbol{\alpha}^t - \boldsymbol{\alpha}^{t+1})^{\mathrm{T}}\boldsymbol{D}^t\boldsymbol{\alpha}^{t+1}\\ =&(\boldsymbol{\alpha}^t - \boldsymbol{\alpha}^{t+1})^{\mathrm{T}}\boldsymbol{X}'^{\mathrm{T}}\boldsymbol{S}(\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{\alpha}^{t+1})\\ =&-(\boldsymbol{e}_s^t - \boldsymbol{e}_s^{t+1})^{\mathrm{T}}\boldsymbol{e}_s^{t+1}.\end{aligned} \tag{33}$$

Combining Lemma 1 and Eq.(33) we have

$$\begin{aligned}&\lambda\sum_{i=1}^{c}\|\boldsymbol{\alpha}_i^t;\mu\|_2^p - \lambda\sum_{i=1}^{c}\|\boldsymbol{\alpha}_i^{t+1};\mu\|_2^p\\ \geq&\lambda\mathrm{tr}((\boldsymbol{\alpha}^{t\mathrm{T}}\boldsymbol{\alpha}^t - \boldsymbol{\alpha}^{(t+1)\mathrm{T}}\boldsymbol{\alpha}^{t+1})\boldsymbol{D}^t)\\ =&\lambda\mathrm{tr}((\boldsymbol{\alpha}^t - \boldsymbol{\alpha}^{t+1})^{\mathrm{T}}(\boldsymbol{\alpha}^t - \boldsymbol{\alpha}^{t+1})\boldsymbol{D}^t) - 2(\boldsymbol{e}_s^t - \boldsymbol{e}_s^{t+1})^{\mathrm{T}}(\boldsymbol{e}_s^{t+1}).\end{aligned} \tag{34}$$

Besides,

$$\begin{aligned}&\|\boldsymbol{s}^{1/2}\odot(\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{\alpha}^t)\|_2^2 - \|\boldsymbol{s}^{1/2}\odot(\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{\alpha}^{t+1})\|_2^2\\ =&\boldsymbol{e}_s^{t\mathrm{T}}\boldsymbol{e}_s^t - \boldsymbol{e}_s^{(t+1)\mathrm{T}}\boldsymbol{e}_s^{t+1}\\ =&(\boldsymbol{e}_s^t - \boldsymbol{e}_s^{t+1})^{\mathrm{T}}(\boldsymbol{e}_s^t - 2\boldsymbol{e}_s^{t+1}) + 2(\boldsymbol{e}_s^t - \boldsymbol{e}_s^{t+1})^{\mathrm{T}}\boldsymbol{e}_s^{t+1}.\end{aligned}$$

This together with (34) leads to

$$\begin{aligned}E_t - E_{t+1} \geq&\lambda\mathrm{tr}((\boldsymbol{\alpha}^t - \boldsymbol{\alpha}^{t+1})^{\mathrm{T}}(\boldsymbol{\alpha}^t - \boldsymbol{\alpha}^{t+1})\boldsymbol{D}^t)\\ &+(\boldsymbol{e}_s^t - \boldsymbol{e}_s^{t+1})^{\mathrm{T}}(\boldsymbol{e}_s^t - \boldsymbol{e}_s^{t+1}) \geq 0\end{aligned} \tag{35}$$

which implies that $E(\boldsymbol{\alpha})$ is non-increasing. Then we have

$$\lambda\sum_{i=1}^{c}\|\boldsymbol{\alpha}_i^t\|_2^p \leq \lambda\sum_{i=1}^{c}\|\boldsymbol{\alpha}_i^t;\mu\|_2^p \leq E(\boldsymbol{\alpha}^t) \leq E(\boldsymbol{\alpha}^1) = N$$

Thus the sequence $\{\boldsymbol{\alpha}^t\}$ is bounded. Moreover, by summing all the inequalities in (35) for all $t\geq$1, we obtain that

$$N = E(\boldsymbol{\alpha}^1) = \sum_{i=1}^{\infty}\lambda\mathrm{tr}((\|\boldsymbol{\alpha}^t - \boldsymbol{\alpha}^{t+1}\|_2^2)\boldsymbol{D}^t) + \|\boldsymbol{e}_s^t - \boldsymbol{e}_s^{t+1}\|_2^2$$

which implies that $\lim_{t\to\infty}\|\boldsymbol{\alpha}^t$-$\boldsymbol{\alpha}^{t+1}\|_2$=0. The proof is completed.

Theorem 4: Any limit point of sequence $\boldsymbol{\alpha}^t$ is a stationary point of Eq.(29). Furthermore, when $p\geq$1, the limit point is globally optimal.

Proof: From Theorem 1, we know that when $p\geq$1, problem (29) is convex. The stationary point must be globally optimal.

Hence we only need to prove sequence $\boldsymbol{\alpha}^t$ converges to a stationary point.

From Theorem 3, the sequence $\{\boldsymbol{\alpha}^t\}$ is bounded, thus there exist a vector $\boldsymbol{\alpha}'$ and another sequence $\{\boldsymbol{\alpha}^{t_j}\}$, such that $\lim_{t\to\infty}\boldsymbol{\alpha}^{t_j}\to\boldsymbol{\alpha}'$. Notice $\boldsymbol{\alpha}^{t_{j+1}}$ solves (30), i.e.,

$$-\boldsymbol{X}'^{\mathrm{T}}\boldsymbol{S}\boldsymbol{y}+\boldsymbol{X}'^{\mathrm{T}}\boldsymbol{S}\boldsymbol{X}'\boldsymbol{\alpha}^{t_{j+1}}+\lambda\boldsymbol{D}^{t_j}\boldsymbol{\alpha}^{t_{j+1}}=\boldsymbol{0} \tag{36}$$

Let $j\to\infty$, (36) implies that $\boldsymbol{\alpha}^{t_{j+1}}$ converges to some $\boldsymbol{\alpha}'^*$, too. From the fact in Theorem 3 that $\lim_{t\to\infty}\|\boldsymbol{\alpha}^t\text{-}\boldsymbol{\alpha}^{t+1}\|_2=0$, we get

$$\|\boldsymbol{\alpha}'-\boldsymbol{\alpha}'^*\|_2=\lim_{t_j\to\infty}\|\boldsymbol{\alpha}^{t_j}-\boldsymbol{\alpha}^{t_{j+1}}\|_2=0$$

That means $\boldsymbol{\alpha}'=\boldsymbol{\alpha}'^*$. Denote $\boldsymbol{\alpha}'$ as $\boldsymbol{\alpha}^*$, when $j\to\infty$ (36) can be rewritten as

$$-\boldsymbol{X}'^{\mathrm{T}}\boldsymbol{S}^*y+\boldsymbol{X}'^{\mathrm{T}}\boldsymbol{S}^*\boldsymbol{X}'\boldsymbol{\alpha}^*+\lambda\boldsymbol{D}^*\boldsymbol{\alpha}^*=\boldsymbol{0}$$

where $\boldsymbol{S}^*$ and $\boldsymbol{D}^*$ are given in Eq.(27) and Eq.(30) with $\boldsymbol{S}^*$ taking place of $\boldsymbol{S}^{t+1}$. Thus $\boldsymbol{S}^*$ satisfies the optimal condition of problem (29).

Note that for expression convenience, we fix $\mu>0$ in aforementioned subsections. However, to make the smoothed IRGSC (19) close to the original problem (18), we reduce the value of μ at every iteration, i.e., $\mu^{t+1}=\mu^t/v$ with $v>1$. It is intuitive to check that our proofs also hold when $\mu^t\to\mu^*>0$.

B. Complexity Analysis

In general, the complexity of IRGSC, RRC [28] and SRC [5] mainly relies on the coding step. It is widely acknowledged that for SRC, the l_1-minimization has a computational complexity of $O(m^2n^{1.5})$ [42], where m is the dimensionality of input samples, and n is the size of training samples. It has been further reported that the mostly used l_1-minimization solvers, e.g., l_1_magic [43] and l_1_ls [44], have an empirical complexity of $O(m^2n^{1.3})$ [46]. For RRC[1], the coding in (8) is also an l_1-norm sparse representation problem, which can be solved via conjugate gradient method [45]. The complexity of RRC is about $O(tk_1k_2nm)$, where k_1 is the iteration number of inner loop and k_2 is the number of iterations to update $\boldsymbol{V}$ [28].

For IRGSC, it has similar optimization process to RRC. We use Eq.(31) to iteratively optimize the representation coefficients, which can be also fulfilled by using conjugate gradient method. Similar to update $\boldsymbol{W}$ and $\boldsymbol{V}$ in RRC, we need to update the feature weights $\boldsymbol{S}$ and the distance weights $\boldsymbol{\Pi}$ in each iteration loop. Therefore, the complexity of IRGSC is also $O(tk_1k_2nm)$. Since that k_1k_2 is with similar order to m by experience, the complexity of Algorithm 1 can be roughly written as $O(tm^2n)$.

Compared to SRC in case of pattern classification without noises, IRGSC usually needs less than 5 iterations to update $\boldsymbol{S}$, its running time is lower than SRC. In pattern classification with noises, IRGSC empirically needs $t\geq 7$. However, the complexity of SRC would be $O(m^2(n+m)^{1.3})$ since it needs to use an extra identity matrix to code the noises. It is clear to say that IRGSC has lower complexity than SRC for pattern classification with noises.

Compared to RRC, IRGSC enjoys more accurate weights learning strategy. This advantage makes IRGSC remove more entries of input features for better performance and achieve optimal representation coefficients within less iterations. Thus the complexity of IRGSC can be further reduced. For example, in FR with real scarf disguise from AR database, RRC uses about 70% pixels and 12 iterations to achieve the accuracy of 90%. However, IRGSC uses only 50% pixels and 7 iterations to achieve similar recognition in the same scenario.

[1]Since RRC with p=1 has better performance than RRC with p=2, in this paper we compare IRGSC to RRC with p=1.

VI. Experimental Analysis

In this section, we perform experiments on benchmark face databases to demonstrate the performance of IRGSC. In Section 6.1, we give the experimental setting of compared approaches and five publically available face databases; in Section 6.2, we evaluate IRGSC for FR under different training size and feature dimension without occlusion; in Section 6.3, we demonstrate the robustness of IRGSC to FR with Gaussian pixel corruption, random block occlusion, mixed noise corruption and real face disguise. Section 6.4 presents the FR results on two unconstrained databases. In Section 6.5, the running efficiency is presented. Finally, some analysis of parameter sensitivity and the behavior of feature weights are given in Section 6.6.

A. Experimental Settings

Five face databases including the AR face [47], the CMU PIE face [48], the Extended Yale B database [49], the LFW [50] and the PubFig datasets [51] are selected to test the effectiveness and robustness of our proposed approach under different scenarios such as illumination, expression changes, and the structural noises led by real disguise, random occlusion, block noise or their mixed noises.

The Aleix Martinez and Robert Benavente (AR) face database contains over 4000 color images featured frontal view faces with various facial expressions, lighting conditions and occlusions corresponding to 126 individuals (70 men and 56 women). The images of 120 subjects were taken in two sessions (separated by two weeks) and each session contains 13 color images. As in [25], a subset that contains 50 male and 50 female subjects was selected in our experiments.

The CMU PIE face image database contains 68 subjects with 41368 face images as a whole. The face images were captured under different pose, illumination and expression. We choose the five near frontal poses (Pose05, Pose07, Pose09, Pose27, Pose29) and use all the images under different illuminations and expressions, thus we get 170 images for each individual. All the face images are manually aligned and cropped to be 32×32 pixels, with 256 gray levels per pixel.

The extended Yale B face database (ExYaleB) contains 38 human subjects under 9 poses and 64 illumination conditions. The 64 images of a subject in a particular pose are acquired at a camera frame rate of 30 frames per second. So there are only small changes in head poses and facial expressions for those 64 images. All frontal-face images marked with P00 are used in our experiments.

The Labeled Faces in the Wild (LFW) contains images of 5,749 different individuals in unconstrained setting. We gathered the subjects including more than 10 samples and then get a dataset with 158 subjects from LFW-a, a revised version of LFW after alignment using commercial software [52]. For each person, 5 samples are randomly selected for training and another 5 samples for test. The images are all cropped and resized to 32×32.

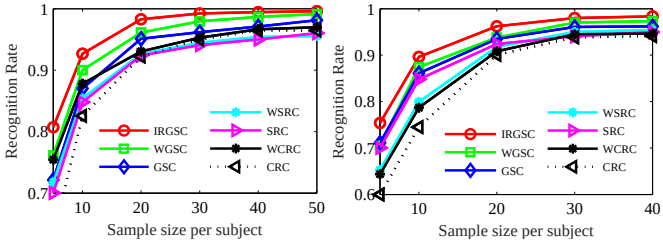
On the PubFig dataset, we follow the same experiment setting as in [53]. We randomly select 20 images for each person and in total 100 subjects are chosen for our experiments, and each image is resized to 64×64 pixels. Ten images for each subject are used as test images and the rest of them are used as training.

The proposed IRGSC is compared to most related existing methods including GSC, WGSC, RRC, SRC, WSRC, CRC, WCRC, NMR and SOC [54]. Note that NMR is a matrix based regression method, which directly uses the gray images as input samples. For all the other vector based approaches in our experiments, each face image is preprocessed as a column vector by connecting its grey intensities in series. Among them, SRC and CRC can be further extended to robust versions, which are denoted as RSRC and RCRC respectively. For IRGSC and RRC, the parameter λ and τ are traversed in $\{$1e-6, 1e-5, 1e-4, 1e-3, 1e-2, 1e-1$\}$ and $\{$0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3$\}$ respectively to report the best result. Without special declarations, the parameter p for IRGSC is set to 1. For SOC, the parameter τ is chosen adaptively, varying from 0.005 to 0.002 with a constant step -0.0005. The parameter β is traversed in $\{$20, 15, 10, 5$\}$ to obtain a best result. For all other comparative methods, the balance parameters λ are also fine-tuned to fit each specific experiment for achieving best performance. The remaining parameters for RRC, SOC, WGSC and RCRC are set to the values suggested by the master copy. All the experiments are implemented by using MATLAB on a 3.10 GHz machine with 3.24GB RAM.

B. Face recognition without Occlusion

We first validate the performance of IRGSC in FR with mixed types of variations such as illumination and expression changes but without occlusion. The ExYaleB database and the PIE database are employed for this purpose. Since there is no corruption, we compare IRGSC with some recently proposed approaches without robustness, such as GSC, WGSC, SRC, WSRC, CRC and WCRC.

1) FR with different samples size: This section tests the effectiveness of the proposed IRGSC under different training size. For the ExYaleB database, we randomly split the database into two parts. One part, which contains n(=10, 20, 30, 40, 50) images for each person, is used as the dictionary, and the other part is used for testing. For each individual in PIE, n(=10, 20, 30, 40) images are randomly selected for training and the rest are used for testing. The averaged recognition rates of 10 runs are plotted against the increasing n in Fig. 2. We can observe that IRGSC achieves the highest recognition rates in all tests. In ExYaleB and PIE, IRGSC has an average improvement of 1.98% and 2.28% respectively over WGSC which ranks second. Besides, the weighted methods WGSC, WSRC, WCRC have a slighted improvement over the non-weighted methods GSC, SRC, CRC, respectively. This demonstrates that data locality is important and an appropriate weight can improve the performance. Furthermore, GSC outperforms SRC and CRC in most cases, which shows the superiority of group sparsity. Overall, our proposed IRGSC which advances both data locality, featured weights, group sparsity and formulates a $l_{2,1}$ mixed norm penalty on the reconstruction coefficients has achieved best performance.



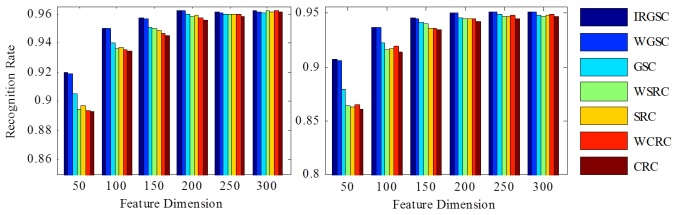
(a)ExYaleB (b)PIE

Fig. 2. Face recognition rate versus number of training samples per subject. (a) ExYaleB (b) PIE.

2) FR with different feature dimension: In this section, we evaluate the performance of the proposed IRGSC under different feature dimension. For both database (ExYaleB and PIE), 20 samples per subject are randomly selected for training and the rest samples are used for testing. The well-acknowledged projection techniques, PCA [3], is used to reduce the dimensionality of original face images. Fig.3 shows the average face recognition accuracy over 10 random runs versus the dimension of the subspace. Much to our surprise, the reduced features which try to get a meaningful low-dimensional representation or reveal the distinctive features of the input data, in fact underperform the original high-dimensional data. From Fig.3, we can conclude that (1) IRGSC achieves better results than the other approaches in all dimensions except that they are slightly worse than WGSC when the dimension is 200 in ExYaleB; (2) compared with other classifiers, IRGSC and WGSC performs much better when the feature dimension is small, for example, the recognition rates of IRGSC, WGSC, GSC, WSRC, SRC, WCRC, and CRC are 90.7%, 90.6%, 87.9%, 86.4%, 86.6%, 86.4%, and 86.1% respectively in PIE with dimension 50. This again verifies the superiority of locality and group sparsity; (3) the recognition rate of IRGSC and WGSC under different projections is very close. The reason for this is that the coding vector solved by Eq.(30) is not accurate enough to estimate $\boldsymbol{s}$ when the feature dimension is very low, which make IRGSC degenerate to WGSC.

C. Face Recognition with occlusion

One of the most interesting advantages of the proposed IRGSC is its robustness with respect to occlusion and noise corruptions. Regarding this issue, the robustness to face occlusion/corruption is achieved by the adaptively learned feature weights. We evaluate the robustness of IRGSC to different types of occlusions in this subsection, such as random pixel corruption, random block occlusion, mixed noise corruption



(a)ExYaleB (b)PIE

Fig. 3. Face recognition rate versus feature dimension with PCA. (a) ExYaleB (b) PIE.

and real face disguise. In the first three experiments of pixel or block corruption, we compare IRGSC methods with RSRC, RCRC, RRC, NMR and WGSC. Among them, RSRC and RCRC are the robust version of SRC and CRC, respectively. Their robustness is achieved by inducing an identity matrix for sparse coding, or equivalently, adopting l_1-norm to measure the coding residual. The robustness of RRC is achieved by assigning iteratively the weights to the pixels according to the logistic function with variable parameters. NMR is a most recently proposed matrix based regression classification method, which not only maintains the structure information of the face images, but also possesses good robustness for the evaluation of the face view. WGSC, even though without robustness to image corruption, has been proved in Section 6.2 to be close to IRGSC under non-occluded scenes and better than other representation based classifiers. In the experiment of real disguise, we further add SOC as a competing method, which achieves significant improvement in real disguise FR problems by employing locality constrained dictionary for occlusion mask estimating and using structured sparsity in the formulation.

1) FR with Pixel Corruption: We randomly select 30 samples per subject of the ExYaleB database for training, and the rest subset for testing. All images are resized to 64×50 pixels. For each test image, a certain percentage of its pixels are randomly replaced by noise uniformly distributed within [0, 255]. One examples of face image with various pixel noises are shown in Fig.4.

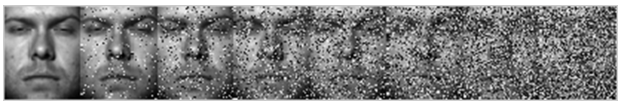


Fig. 4. Face samples with different percentage of pixel corruption (from 0% to 70%)

Fig.5 exhibits a representative recognition example of the 5th subject with 60% random pixel corruption. Fig.5(a) is the original face image, and Fig.5(b) shows the corrupted face images with 60% pixel occlusion, which is difficult to recognize, even for human eyes. Fig.5 (c), (d), (e), (f), (g) and (h) shows the reconstructed images of IRGSC, RRC, RSRC, RCRC, WGSC and NMR respectively. It can be seen that the reconstructed face image of IRGSC is more faithful to the original image than other methods. Furthermore, the visual quality (especially for classification) of Fig.5(c) is better than Fig.5(a) since the illumination changes which bring difficulties to recognition has alleviated. For RRC, the shadows which exist in Fig.5(d) have not been removed. For RSRC, the reconstruction image in Fig.5(e) is more like an average image, which looks as another person. For RCRC and WGSC, we only can see two contours of human face from their recovered images since their non-sparseness and non-robustness, respectively. The reconstructed image in Fig. 6 (h) is only a hallucination face since that NMR is good at structure information exploring, but poor at Gaussian noises, which make it less discriminative and with poorer classification performance (shown in Table I).

(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 5. Recognition under random pixel corruption. (a) Original test image. (b) Test image with 60% random corruption. (c) Reconstructed images of IRGSC, (d) RRC, (e) RSRC, (f) RCRC, (g) WGSC, (h) NMR.

Table 1 lists the average accuracy of IRGSC, RRC, RSRC, RCRC, WGSC and NMR against the noise level over 10 runs. It demonstrates that the proposed IRGSC consistently outperforms other algorithms for all levels of corruption. Particularly, WGSC is sensitive to the outliers, with accordingly lower accuracy than others. RCRC performs about the same as RSRC. Their differences of accuracy are between 0.1 and 0.7 percent. RRC ranks two in most cases except that it is worse than RCRC and RSRC when the corruption reaches over 60%. Overall, the gains of IRGSC over the second accuracy is 0.4%, 1.2%, 1.2%, 0.9%, 1.4%, 3.8%, and 0.6% along with the increasing of corruption.

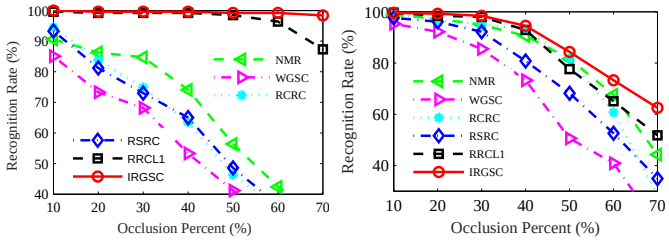
TABLE I
RECOGNITION RATES (%) OF IRGSC, RRC, RSRC, RCRC, WGSC AND NMR VERSUS DIFFERENT PERCENTAGE OF PIXEL CORRUPTION

Corruption (%)	10	20	30	40	50	60	70
NMR	93.7	90.2	85.1	76.5	67.1	51.8	29.8
WGSC	94.1	94.1	90.9	83.9	72.2	48.2	29.1
RCRC	97.0	96.6	96.5	95.3	94.6	91.8	83.9
RSRC	96.5	96.2	95.8	95.2	94.8	91.9	84.5
RRC	99.6	98.8	98.0	97.1	96.5	91.0	82.0
IRGSC	100	100	99.2	98.4	97.4	95.7	85.1

2) FR with Block Occlusion: In this subsection, we design three block occlusion experiments using the datasets being identical to last section. In the first experiment, we replace the 10∼70% pixels of each testing image using a white or black block. The location of the square block is randomly selected. Fig.6 shows some samples of blocked face images from the ExYaleB database with different level of occlusions. Fig.7 plots recognition rates of NMR, WGSC, RCRC, RSRC, RRC and IRGSC under different levels of occlusions (from 10% to 70%). With the increment of block occlusion, we observe that IRGSC significantly outperforms the other methods. When the black occlusion percentage is 70%, the recognition rate of

IRGSC is 10.7%, 18.2%, 27.6%, 33.9% and 45.7% higher than RRC, NMR, RSRC, RCRC and WGSC, respectively. In the case of white block occlusion, RRC achieves comparable results when the occlusion percentage is lower than 50%. However, the recognition rate of IRGSC is 2.8% and 11.3% higher than that of RRC when the occlusion percentage is 60% and 70% respectively. NMR ranks third in both the black and the white occlusion cases. It uses nuclear norm for better structure learning, but lacks feature learning mechanism. When we resize the experimental data to 96×84 for more structural information, NMR achieves comparable results with ours under 50% black block occlusion. However, it needs more memory size and still lags behind IRGSC by 4.7% and 12.7% when the occlusion rates are 60% and 70%, respectively. This further demonstrates the superiority of our adaptive feature learning mechanism.

Fig. 6. The face images with white block occlusion (from 10% to 70%). (a) Test images with white block occlusion. (b) Test images with black block occlusion.



(a) White block occlusion (b) Black block occlusion

Fig. 7. The recognition rates (%) of NMR, WGSC, RCRC, RSRC, RRC and IRGSC with the block occlusion percentage ranging from 10% to 70%

In the second experiment, we replace the 10∼60% pixels of each testing image using another randomly selected face image. Fig.8 shows the samples of testing images, from which we can see that the occlusion of a certain size is located on the random position which is unknown to the FR algorithms. It is relatively difficult than the first experiment, since the pixels of occlusion area are close to the effective coverage. Table 2 lists the detailed recognition rates of IRGSC, RRC, RSRC, RCRC and WGSC and NMR under the occlusion percentage from 10% to 60%. We can see that the obstruction of another face images may produce some unexpected results. When the block occlusion ratio is 10%, RCRC and RSRC, which try to solve the problem of robustness, in fact underperform WGSC that is without robustness. When the block occlusion ratio is higher than 30%, RCRC outperforms RRC that is claimed to be more robust in [26]. A probable reason is that RRC mainly addresses the occlusion problem with big changes of pixel values. The performance of NMR is better than WGSC under all the levels of face block occlution, and it outperforms RRC when the block occlusion ratio is 40% and 60%. Nevertheless, the proposed IRGSC consistently and visibly performs the best for all levels of face occlusions.

TABLE II
RECOGNITION RATES (%) OF IRGSC, RRC, RSRC, RCRC, WGSC AND NMR UNDER DIFFERENT LEVELS OF FACE BLOCK OCCLUSION

Corruption (%)	10	20	30	40	50	60
NMR	99.2	96.9	93.7	89.1	78.1	65.5
WGSC	97.6	96.1	90.2	80.0	67.8	51.0
RCRC	97.3	97.3	95.7	89.5	83.1	68.2
RSRC	96.9	96.9	94.5	88.6	72.9	60.4
RRC	99.6	98.4	94.1	88.6	79.2	64.7
IRGSC	99.6	98.8	95.3	90.4	84.5	69.4

Fig. 8. The face images with randomly selected face occlusion (from 0% to 60%).

In the third experiment, we test the effectiveness of IRGSC on dataset whose face images have mixed types of corruption. Different portions of testing images, from 10 to 50 percent, are simultaneously occluded by mixed noise corruption (a randomly selected face image together with random pixel corruption) at random locations as shown in Fig.9. The experimental results are shown in Table 3. Obviously, it is the most challenging experiment in this subsection. We can see that the recognition rates are clearly lower than the last two experiments. Encouragingly, the proposed algorithm still outperforms other methods in all corruption levels except that it is equal to RRC when the occlusion percent is 10 and 30%. Especially, when the occlusion percent is 20%, it has an improvement of 0.4%, 4.3%, 3.6% and 17.3% and 17.4% over RRC, RSRC, RCRC, WGSC and NMR, respectively.

Fig. 9. The face images with composite noise (from 0% to 50%).

TABLE III
RECOGNITION RATES (%) OF IRGSC, RRC, RSRC, RCRC, WGSC AND NMR UNDER DIFFERENT LEVELS OF MIXED NOISE CORRUPTION

Corruption (%)	10	20	30	40	50
NMR	92.0	78.7	59.2	39.6	19.2
WGSC	93.7	78.8	57.3	35.7	12.9
RCRC	96.5	92.5	90.2	74.9	34.5
RSRC	95.7	91.8	87.1	63.9	22.7
RRC	99.2	95.7	91.4	75.3	45.1
IRGSC	99.2	96.1	91.4	76.7	47.1

3) FR with Real Disguise: After experimenting under random pixel corruption and block occlusion scenarios, we further

TABLE V
RECOGNITION RATES (%) BY COMPETING METHODS ON THE LFW AND PUBFIG DATABASES

Databases	CRC	WCRC	SRC	WSRC	GSC	WGSC	NMR	RCRC	RSRC	RRC	IRGSC
LFW	44.7	44.8	40.0	41.2	47.5	47.6	48.2	45.3	42.8	53.2	56.3
PubFig	35.3	35.3	45.0	36.7	37.1	37.5	44.5	43.6	47.0	42.2	48.5

test different approaches in coping with real possible disguise occlusion. In the first test, the images were resized to 64×50. Fig.10 illustrates the classification process of IRGSC, SOC, RRC, NMR, RSRC, RCRC and WGSC by using an example. Fig.10(a) shows a test image with scarf; Fig.10 (b), (c), (d), (e), (f), (g) and (h) show the reconstructed images of IRGSC, SOC, RRC, NMR, RSRC, RCRC and WGSC, respectively; Fig. 10(i) displays the representation coefficients of all computing methods with the correct class marked as red. Fig. 10(j) displays the residuals per class of all computing methods with the lowest value marked as red. From Fig. 10(b) to (h), we see that RSRC, RCRC and WGSC fail to regain a clear human face. For RSRC in particular, the reconstructed image looks like a man cultivated a huge beard and thick mustache, which will seriously infect the classification results. The remaining four methods all successfully regain a clear face. However, the image recovered by IRGSC is more coherent to the original input face image. From Fig.10 (i), it can be seen that for IRGSC the dictionary atoms with the same label as the testing sample have clearly bigger coefficients compared to other 6 competing methods. The residuals in Fig. 10(j) verify that only IRGSC classifies the test image into the right class, but the other 6 methods all fail.

In the second test, we use the similar experiment setting as in [28] to conduct FR in real disguise with variations of illumination and longer data acquisition interval. 400 images (4 neutral images with different illuminations per subject) of non-occluded frontal views in Session 1 were used for training, while the disguised images (3 images with various illuminations and sunglasses or scarves per subject per Session) in Sessions 1 and 2 for testing. Table 4 lists the results by competing methods. Clearly, the IRGSC methods achieve much better results than SOC, RRC, RSRC, RCRC, WGSC and NMR. Interestingly, SOC that try to solve the problems of RRC, in fact underperform RRC. NMR also performs poorly in this test, since the limited training quantities and image resolution contain limited structural information. Overall, the average improvements of IRGSC over RRC, SOC, RSRC, RCRC, NMR and WGSC are respectively 0.35%, 11.7%, 21.2%, 31.2%, 40.7% and 45.4% on sunglasses, and respectively 4.6%, 14.2%, 66.9%, 33.2%, 30.6% and 39.9% on scarf.

D. Face recognition with unconstrained setting

As we know, the faces in LFW and PubFig were acquired under unconstrained setting and inaccurate alignment, which makes them extremely challenging for FR. Some samples for a person from LFW and PubFig are showed in Fig. 11a and Fig. 11b, respectively. Table 5 contains the results of all competing methods for FR in these two datasets. SOC is not included in this experiment since it is designed exclusively for FR

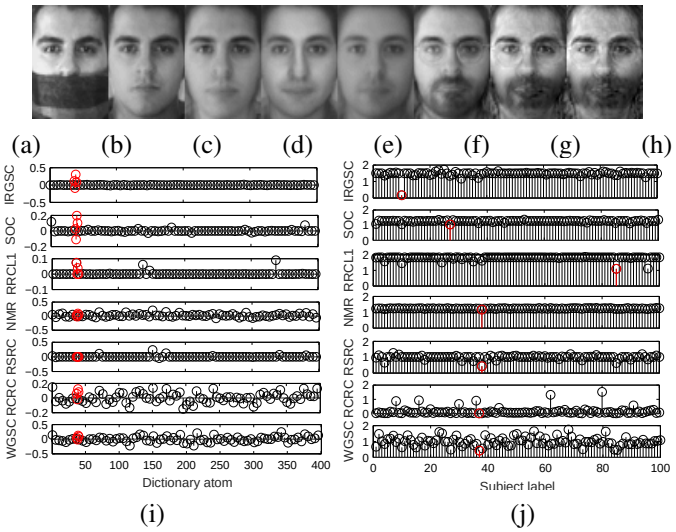


Fig. 10. Recognition with scarf disguise. (a) Original test image from AR. (b) Reconstructed images of IRGSC, (c) SOC, (d) RRC, (e) NMR, (f) RSRC, (g) WGSC. (h) WGSC, (i) Coefficients $\boldsymbol{\theta}$, (j) Residuals of each class

Fig. 11. Some face images for one person from (a) LFW database and (b) PubFig database.

with block occlusion, but its performance is not competitive in general cases without occlusion. For LFW, it is evident that IRGSC obtains the highest recognition rate: 56.3%. RRC and NMR achieve the second and third highest recognition rate: 53.2% and 48.2%, respectively. For PubFig, our method also achieves the better performance than all the competing approaches. Although RRC obtains relatively competitive recognition rates for LFW, it lags behind IRGSC by 6.3%

TABLE IV
RECOGNITION RATES (%) BY COMPETING METHODS ON THE AR DATABASE WITH REAL DISGUISE OCCLUSION

Classifiers	Session 1		Session 2	
	Sunglasses	Scarves	Sunglasses	Scarves
NMR	72.3	72.3	35.3	45.3
WGSC	66.3	62.7	32.0	36.3
RCRC	80.3	70.3	46.3	42.0
RSRC	89.3	32.3	57.3	12.7
RRC	99.0	93.3	89.3	76.3
SOC	95.2	88.5	70.4	61.9
IRGSC	99.0	96.7	90.0	82.0

for PubFig. This further demonstrate that it is reasonable to learn the feature weights adaptively and adopt $l_{2,p}$-norm as regularized constraint.

All the above experimental results of our method outperform other compared approaches. This confirms that IRGSC is robust to real disguise, mixed occlusion, illumination and unrestricted environment. However, the accuracy of our method in LFW and PugFig is merely around 50% which lags far behind human-level performance. To improve performance, we further introduce a learned convolutional neural network (CNN) [55] as a feature extractor to train our input samples. By using PCA as the dimensionality reduction tool, the recognition rate of IRGSC reaches 98.4% and 97.2% in LFW and PugFig with 300 dimensions, respectively. These results are competitive to state-of-the-art methods such as DeepFace [56]. Since IRGSC has much simpler theoretic mechanism and can be applied in mixed types of corruption in limited quantities of training samples, it has extensive application prospects.

E. Running Time

Apart from accuracy, computational cost is another important issue for different classifiers. We test the running time of 5 robust vector based methods in this section, including RCRC, RSRC, RRC, SOC and IRGSC, is assessed in practical FR experiments with mixed noise corruption and with real disguise. For SRC, we adopt $l_1\_l_s$ to implement the sparse coding step. For RCRC, RSRC and RRC, we use the codes implemented by the authors. For IRGSC, we implement it in three different versions, i.e., IRGSC without feature weights (we name it as IRGSC_WF for abbreviation), IRGSC without distance weights (we name it as IRGSC_WD for abbreviation), and IRGSC. Note that IRGSC_WF, which degraded from IRGSC, is actually same as WGSC. They share the same accuracy and running time.

The first experiment conducted on the ExYaleB database with 20% mixed noise corruption. The settings are the same as Section 6.3.2. Table 6 lists the average running time of 10 runs. The recognition rate of IRGSC_WD is 96.0%. The recognition rate of other methods can be found in Table 3. As shown in Table 3 and Table 6, IRGSC_WF is the fastest algorithm among all the competing methods, however, its recognition rate is the lowest. RSRC implemented by SPAM is the slowest method with similar recognition rate to the second fastest method RCRC. Compared to RRC that ranks two in recognition rate, IRGSC has higher recognition rate with comparable computation expense. IRGSC_WD, which is a degradation of IRGSC, has much less computational cost and simultaneously has higher recognition rate than RRC.

The second experiment is FR on the AR database with real disguise. The experimental settings are same as Section 6.3. The average recognition rates and computational time of competing methods are reported in Table 7. Clearly, IRGSC_WF has the least computation expense, followed by RCRC. RSRC and SOC have rather high computation burden since they both utilize an additional matrix to code occlusion. For the recognition rate, SOC's performance is medium among all competing methods. Considering both the recognition rate and running cost, IRGSC_WD is better than RRC encouragingly. IRGSC gets the highest recognition rates in all cases with sacrifice of a little computational cost.

TABLE VII
AVERAGE RECOGNITION RATE (%) AND RUNNING TIME (SECONDS) OF COMPETING METHODS ON AR DATABASE WITH REAL FACE DISGUISE.

Classifiers	Recognition rate	Running time
RCRC	59.7	0.17
RSRC	47.9	3.52
RRC	89.5	1.08
SOC	79.0	3.54
IRGSC_WF	49.3	0.05
IRGSC_WD	90.5	0.75
IRGSC	91.9	1.32

F. The behavior of parameters and feature weights

We discuss the influence of parameter λ and l in IRGSC on the recognition performance in this section. As described in Section 4, the parameter l is a pivotal parameter to discriminate inliers or outliers. Our experiments adopts l=$m\tau$, $\tau\in(0,\ 1)$, to estimate l. Consequently, it is indispensable to debate the selection of τ. The parameter λ balances the contribution of the reconstruction error and the regularization term, which is universally used in all the mentioned regression representation classification approaches. We conduct the experiment with various level of black block occlusion (experimental settings are same as Section 6.3.2) as an example. Fig.12 plots the recognition rates of IRGSC versus λ and τ in roughly estimating l under 20%, 40%, and 60% block occlusion. It can be obtained from Fig. 12 that the performance of our method is relatively less sensitive to the balance parameter λ compared to τ. Moreover, there is a regular and unimodal trend for different τ under the settled occlusion percentage. These properties make the final value of λ easy to be determined. On the other hand, IRGSC could get outstanding performance (i.e., $\geq$96%) in a wide span of τ for mild corruption (i.e., 20%). For all percentages of block occlusion, the best τ gets lower with the increasing of occlusion level. It is reasonable because more features could be trusted when there are smaller percent of outliers. It makes our proposed IRGSC method be conveniently tuned to various range of noise corruption.

In Algorithm 1, it has been proved empirically and theoretically that the value of p will influence the sparsity of the coding coefficients [12], [39]. In this subsection, we evaluate the impact of p on classification performance. The experiment is conducted on the PIE database, whose experimental setting is the same as that in Section 6.1 but with different values of p. Fig.13 shows the performance comparison of IRGSC with different p. It can be seen that the accuracy curves fluctuate slightly with the changes of p, which means that a carefully chosen p will further improve the performance of IRGSC. With the increase of training samples, the optimal value of p moves from 0.1 to 1. The most possible optimum ranges from 0.5 to 0.8. These observations verify the theoretical statements and give us clear instructions for model construction.

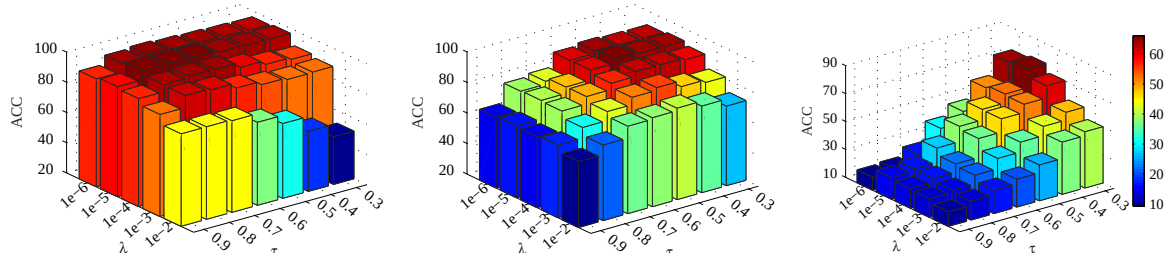


Fig. 12. Recognition performance versus λ and τ in ExYaleB under (a) 20% occlusion, (b) 40% occlusion, and (c) 60% occlusion

TABLE VI
RUNNING TIME (SECONDS) OF COMPETING ALGORITHMS ON EXYALEB DATABASE WITH MIXED NOISE CORRUPTION

Classifiers	RCRC	RSRC	RRC	IRGSC_WF	IRGSC_WD	IRGSC
Running time per test image	1.5	14.8	12.7	0.6	9.6	13.2

It is necessary to check whether our idea of weights learning mechanism works as expected, i.e. whether the weights $\boldsymbol{s}$ adjusted adaptively as the residue changes under indeterminate condition. Fig. 14 illustrates the learned feature weights $\boldsymbol{s}$ under (a) pixel corruption, (b) white block occlusion, (c) black block occlusion, (d) image occlusion, (e) mixed corruption under 50% percentage coverage rate and the real disguise of (f) sunglass and (g) scarf. The upper row of Fig. 14 is grouped of different query samples and the corresponding lower one is the final feature weights. From Fig. 14 we could see that IRGSC assigns bigger weights (near white region) to the un-occluded pixels, and assigns lower weights (near black region) to the occluded pixels. Especially for the block occlusion environment, it is particularly clear that the corresponding occluded region is assigned to values close to 0 for removal. Moreover, IGRSC correctly classifies the region of shadow into useless features. All these results further verify the robustness of our methods.

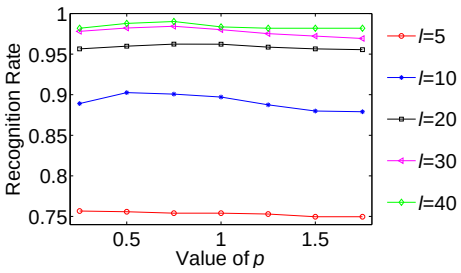
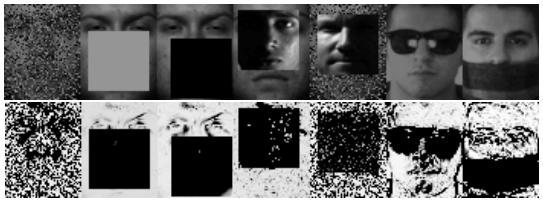


Fig. 13. Recognition performance versus p under different training size.



(a) (b) (c) (d) (e) (f) (g)

Fig. 14. The behavior of learned weights under different corruption and real disguise.

the query samples. By combining the feature weights learning and distance weights learning into one objective function, we can effectively remove the outlier samples as well as outlier features. For achieving the optimal representation coefficients, we first smooth the objective function by introducing regularization terms. Then an adaptive and iterative strategy is applied for solving the relaxed problem, we further provide a general proof to show that the solution by IRGSC is a stationary point (globally optimal if the problem is convex). The proposed IRGSC method is extensively evaluated on FR with different scenarios, including variations of illumination, expression, occlusion, and corruption. The experimental results clearly demonstrated that IRGSC outperforms significantly previous state-of-the-art methods, such as RRC, WGSC and SRC. In particular, two trimming of IRGSC, IRGSC_WF and IRGSC_WD, could also achieve very high accuracy in certain scenarios but with lower computational cost, which makes them a very preferable candidate for practical FR systems. An interesting future work is to transform IRGSC into tensor or kernel variants to make it be qualified for more complex applications.

VII. CONCLUSION

This paper presents a novel iterative re-constrained group spare classification (IRGSC) method for robust FR. An effective optimization strategy is also proposed which facilitates the implementation of IRGSC. One important advantage of IRGSC over the existing algorithms is its robustness to mixed types of noises (e.g., occlusion, corruption, etc.,) by seeking for an efficient feature selection mechanism to precisely regress

REFERENCES

[1] L. Q. Liu, X. Chao, H. W. Zhang, Z. H. Niu, M. Yang, and S. C. Yan, "Deep aging face verification with large gaps," *IEEE Trans. Multimedia*, vol. 18, no. 1, 2016, pp. 64-75.
[2] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, 2016, pp. 188-194.
[3] X. D. Jiang, "Asymmetric principal component and discriminant analyses for pattern classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 5, May 2009, pp. 931-937.

[4] I. Naseem, R. Togneri, and M. Bennamoun, "Linear regression for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, Nov 2010, pp. 2106-2112.

[5] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, Feb. 2009, pp. 210-227.

[6] J. Qian, L. Luo, J. Yang, F. L. Zhang, and Z. C. Lin, "Robust nuclear norm regularized regression for face recognition with occlusion," *Pattern Recognition*, vol. 48, no. 10, Oct 2015, pp. 3145-3159.

[7] L. Zhang, M. Yang, and X. C. Feng, "Sparse representation or collaborative representation: Which helps face recognition?" in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 471-478.

[8] J. Yang, L. Luo, J. J. Qian, Y. Tai, F. L. Zhang, and Y Xu, Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 156-171, Jan 2017.

[9] D. X. Zhong, Z. C. Xie, Y. R. Li, and J. Q. Han, "Loose L1/2 regularised sparse representation for face recognition," IET Computer Vision, vol. 9, no. 2, Apr 2015, pp. 251-258.

[10] J. Huang, F. P. Nie, H. Huang, and Chris. D, "Supervised and Projected Sparse Coding for Image Classification," in *Proc. 27th AAAI Conf. Artificial Intelligence*, 2013, pp. 438-444.

[11] X. D. Jiang, J. Lai. "Sparse and dense hybrid representation via dictionary decomposition for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 5, May. 2015, pp. 1067-1079.

[12] J. Wu, R. Timofte, and L. Van Gool. "Learned collaborative representations for image classification," In *IEEE Winter Conf. Applications of Computer Vision*, Jan 2015, pp. 1-8.

[13] U. Srinivas, Y. M. Suo, M. Dao, V. Monga, and T. D. Tran, "Structured Sparse Priors for Image Classification," *IEEE Trans. Image Process.*, vol. 24, no. 6, Jun 2015, pp. 1763-1776.

[14] S. Bahrampour, N. M. Nasrabadi, A. Ray, and W. K. Jenkins, "Multimodal task-driven dictionary learning for image classification," *IEEE Trans. Image Process.*, vol. 25, no. 1, Jan 2016. pp. 24-38.

[15] Q. Qiu, R. Chellappa, "Compositional dictionaries for domain adaptive face recognition," *IEEE Trans. Image Process.*, vol. 24, no. 12, Dec 2015, pp. 5152-5165.

[16] S. Q. Zhang, X. M. Zhao, "Locality-sensitive kernel sparse representation classification for face recognition," *J. Vis. Commun. Image R.*, vol. 25, no. 8, Nov 2014, pp. 1878-1885.

[17] C. Y. Lu, H. Min, J. Gui, L. Zhu, and Y. K. Lei, "Face recognition via weighted sparse representation," *J. Vis. Commun. Image R.*, vol. 24, no. 2, Feb 2013, pp. 111-116.

[18] Z. Fan, M. Ni, Q. Zhu, and E. Liu, "Weighted sparse representation for face recognition," *Neurocomputing*, vol. 151, no. 3, Mar 2015, pp. 304-309.

[19] R. Timofte, L. V. Gool, "Adaptive and weighted collaborative representations for image classification," *Pattern Recognition Letters*, vol. 43, no. 7, Jul 2014, pp. 127-135.

[20] Y. Chao, Y. Yeh, Y. Chen, Y. Lee, and Y. Wang, "Locality-constrained group sparse representation for robust face recognition," in *Proc. IEEE Int. Conf. Image Processing*, 2011, pp. 761-764.

[21] X. Tang, G. C. Feng, and J. X. Cai, "Weighted group sparse representation for undersampled face recognition," *Neurocomputing*, vol. 145, no. 12, Dec 2014, pp. 402-415.

[22] I. Naseem, R. Togneri, and M. Bennamoun, "Robust regression for face recognition," *Pattern Recognition*, vol. 45, no. 1, Jan 2012, pp. 104-118.

[23] L. Zhang, M. Yang, X. Feng, Y. Ma, and D. Zhang, "Collaborative Representation based Classification for Face Recognition," Comput., Hong Kong Polytechnic University, Hong Kong, Tech. Rep. 2014.

[24] R. He, W. S. Zheng, T. Tan, and Z. Sun, "Maximum correntropy criterion for robust face representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, Aug 2011. pp. 1561-1576.

[25] R. He, W. S. Zheng, T. Tan, and Z. Sun, "Half-quadratic based iterative minimization for robust sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, Feb 2014, pp. 261-275.

[26] L. Luo, J. Yang, J. J. Qian, and Y. Tai, Nuclear-l1 norm joint regression for face reconstruction and recognition with mixed noise, *Pattern Recognition*, vol. 48, no. 12, pp. 3811-3824, DEC 2015.

[27] J. H. Chen, J. Yang, L. Luo, J. J. Qian, and W. Xu, Matrix variate distribution-induced sparse representation for robust image classification, *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 10, pp. 2291-2300, Oct 2015.

[28] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Regularized robust coding for face recognition," *IEEE Trans. Image Process.*, vol. 22, no. 5, May 2013, pp. 1753-1766.

[29] J.J. Qian, J. Yang, General regression and representation model for face recognition, in *Proc. IEEE Conf. Comput. Vis. Pattern. Recognit.*, Jun. 2013, pp. 166C172.

[30] M. S. Cui, S. Prasad, "Class-dependent sparse representation classifier for robust hyperspectral image classification," *IEEE Trans. Geoscience and Remote Sensing*, vol. 53, no. 5, May 2015, pp. 2683-2695.

[31] J. Jiang, R. Hu, Z. Wang, and Z. Han, "Noise robust face hallucination via locality- constrained representation," *IEEE Trans. Multimedia*, vol. 16, no. 5, May 2014, pp. 1268-1281.

[32] W. Y. Liu, Z. D. Yu, L. J. Lu, Y. D. Wen, H. Li, and Y. X. Zou, "KCRC-LCD: Discriminative kernel collaborative representation with locality constrained dictionary for visual categorization," *Pattern Recognition*, vol. 48, no. 10, Oct 2015, pp. 3076-3092.

[33] J. L. Jiang, L. Zhang, and J. Yang, "Mixed noise removal by weighted encoding with sparse nonlocal regularization," *IEEE Trans. Image Process.*, vol. 23, no. 6, Jun 2014, pp. 2651-2662.

[34] C. L. P. Chen, L. C. Liu, L. Chen, Y. Y. Tang, and Y. C. Zhou, "Weighted couple sparse representation with classified regularization forimpulse noise removal," *IEEE Trans. Image Process.*, vol. 24, no. 11, Nov 2015, pp. 4026-4041.

[35] H. Yan, J. Yang, "Locality preserving score for joint feature weights learning," *Neural Networks*, vol. 69, no. 9, Sep 2015, pp. 126-134.

[36] Z. C. Li, J. Liu, J. H. Tang, and H. Q. Lu, "Robust structured subspace learning for data representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 10, Oct 2015, pp. 2085-2098.

[37] F. P. Nie, X. Q. Wang, and H. Huang, "Clustering and projected clustering with adaptive neighbors," *in Proc. Int. Conf. Machine Learning*, 2014, pp. 977-986.

[38] S. Boyd, L. Vandenberghe, "Convex optimization," Cambridge, UK: Cambridge University Press, 2004.

[39] C. Y. Lu, Z. C. Lin, and S. C. Yan, "Smoothed low rank and sparse matrix recovery by iteratively reweighted least squares minimization," *IEEE Trans. Image Process.*, vol. 24, no. 2, Feb 2015, pp. 646-654.

[40] I. Daubechies, R. Devore, M. Fornasier, and C. S. Gunturk, "Iteratively reweighted least squares minimization for sparse recovery," *Commun. Pure Appl. Math.*, vol. 63, no. 1, 2010, pp. 1-38.

[41] C. Chen, J. Z. Huang, L. He, and H. S. Li, "Preconditioning for accelerated iteratively reweighted least squares in structured sparsity reconstruction," in *Proc. IEEE Conf. Comput. Vis. Pattern. Recognit.*, Jun. 2014, pp. 23-28.

[42] Y. Wang, J. J. Wang, Z. B. "Restricted p-isometry properties of nonconvex block-sparse compressed sensing," Signal Processing, vol. 104, no. 11, Nov 2014, pp. 188-196.

[43] Y. Nesterov, A. Nemirovskii, "Interior-point polynomial algorithms in convex programming," Philadelphia, PA: SIAM, 1994.

[44] E. Cands and J. Romberg. (2005). L1 -Magic: Recovery of Sparse Signals via Convex Programming [Online]. Available: http://www.acm.caltech.edu/l1magic.

[45] S. J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "A interiorpoint method for large-scale 1 -regularized least squares," *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 4, Sep 2007, pp. 606-617.

[46] J. R. Shewchuk, "An introduction to the conjugate gradient method without the agonizing pain," School Comput. Sci., Carnegie Mellon Univ., Pittsburgh, PA, Tech. Rep. CMU-CS-94-125, 1994.

[47] A. Martinez, R. Benavente, "The AR face database," Centre de Visi per Computador, Universitat Autnoma de Barcelona, Bellaterra, Barcelona, Tech. Rep. 24, 1998.

[48] A. Georghiades, P. Belhumeur, and D. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, Jun. 2001, pp. 643-660.

[49] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image and Vis. Comput.*, vol. 28, no. 5, 2010, pp. 807-813.

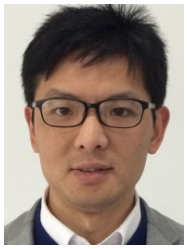
[50] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, Labeled faces in the wild: A database for studying face recognition in unconstrained environments, Univ. Massachusetts, Amherst, MA, USA, Tech. Rep. 07-49, 2007.

[51] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, Attribute and simile classifiers for face verification, in *Proc. ICCV*, 2009, pp. 365-372.

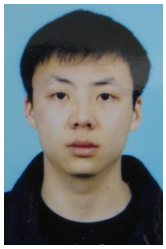
[52] P. Zhu, L. Zhang, Q. Hu, and S. C. K. Shiu, Multi-scale patch based collaborative representation for face recognition with margin distribution optimization, in *Proc. Euro. Conf. Comput. Vis.*, Oct. 2012, 822-835

[53] L. Luo, L. Chen, J. Yang, J. J. Qian, B. Zhang, Tree-structured nuclear norm approximation with applications to robust face recognition, *IEEE Trans. Image Process.*, vol. 25, no. 12, pp. 5757-5767, Dec. 2016.

[54] Y. D. Wen, W. Y. Liu, M. Yang, Y. L. Fu, Y. J. Xiang, and R. Hui, "Structured Occlusion Coding for Robust Face Recognition," *Neurocomputing*, vol. 178, no. 2, Feb 2016, pp. 11-24.
[55] O. M. Parkhi, A. Vedaldi, A. Zisserman, Deep face recognition, in *Proc. Br. Mach. Vis. Conf.*, 2015, pp. 41.1-41.12.
[56] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, Deep-Face: Closing the gap to human-level performance in face verification, in *Proc. IEEE Conf. Comput. Vis. Pattern. Recognit.*, Jun. 2014, pp. 1701-1708.



Jianwei Zheng received the B.S. degree in electronic and computer engineering and the Ph.D. degree in control theory and control engineering from Zhejiang University of Technology, in 2005 and 2010, respectively. He is currently an associate professor with the college of Computer Science and Technology, Zhejiang University of Technology. His research interests include machine learning and compressive sensing. He has authored over 40 journal and conference papers in these areas.



Ping Yang born in 1992. Reads a Ph.D. degree in control theory and control engineering at Zhejiang University of Technology. His main research interests include machine learning and computer vision.

Shengyong Chen received the Ph.D. degree in computer vision from City University of Hong Kong, Hong Kong, in 2003. He is currently a Professor of Tianjin University of Technology and Zhejiang University of Technology, China. He received a fellowship from the Alexander von Humboldt Foundation of Germany and worked at University of Hamburg in 2006-2007. His research interests include computer vision, robotics, and image analysis. Dr. Chen is a Fellow of IET and senior member of IEEE and CCF. He has published over 100 scientific papers in international journals. He received the National Outstanding Youth Foundation Award of China in 2013.

Guojiang Shen received the B.S. degree in Control Theory and Control Engineering and the Ph.D. degree in Control Science and Engineering from Zhejiang University, Hangzhou, China, in 2004. His research interests include machine learning and intelligent system.

Wanliang Wang received the Ph.D. degree in control theory and control engineering from Tongji University, Shanghai, China, in 2001. He is currently a full Professor of Zhejiang University of Technology, China. In 2002, he visited the University of Manchester Institute of Science and Technology, and also visited the Georgia Institute of Technology, and the University of Michigan, USA. His current research interests include artificial intelligence. He was a recipient of the National Outstanding Teacher Award in 2008, and the First National Teacher of Ten Thousand Plan Award in 2014.